On the Anatomy of Cyberattacks[*]

Jin-Wook Chang Kartik Jayachandran Carlos A. Ramírez

Ali Tintera

February 26, 2024

Abstract

Using detailed information on cyberattacks and establishments in the United States, we study whether and how an establishment's characteristics can alter the likelihood of cyberattacks. We find that larger establishments and establishments of publicly traded companies are more likely targets.

JEL Codes: D39, D22, M15.

Keywords: cyberattacks, cybersecurity, hacks, cyber risk.

Highlights:

- Larger establishments are more likely targets of cyberattacks.
- Establishments of publicly traded companies are also more likely targets.
- Results are robust to variation on merging methodologies and regression specifications.

[*] All authors are with the Federal Reserve Board. The views expressed herein are ours and might not necessarily reflect those from Federal Reserve Board or other members of its staff. Emails: jin-wook.chang@frb.gov, kartik.jayachandran@frb.gov, carlos.ramirez@frb.gov [corresponding author], and ali.tintera@frb.gov. There are no competing interests to declare.

1 Introduction

The increased frequency and severity of cyberattacks has recently attracted considerable attention. Despite that cybersecurity threats consistently ranked among the top 10 concerns of business, government, and academic leaders, the distribution of cyberattacks across institutions is, at best, imperfectly understood, as public data are scant and mostly anecdotal.[1] This paper partially fills this gap by using detailed information on cyberattacks and establishments in the U.S. and studying which types of establishments are more likely to be targeted.

Ex-ante, it is not obvious which institutions are more likely to become victims of cyberattacks. This is because, from a theoretical perspective, the equilibrium distribution of cyberattacks depends on hackers' motivation and institutions' response, both of which are difficult to observe.[2] To understand this observation more clearly, consider a simple economy wherein hackers are purely financially motivated. And assume hackers target larger institutions with the expectation of obtaining higher ransoms. Considering this strategy, larger institutions would increase their investments in cybersecurity, making it more difficult to implement successful attacks. Thus, on expectation, targeting larger institutions may become less profitable. A similar idea applies to smaller institutions. Here, however, expected profits from successful attacks might be smaller, as smaller institutions might be unable to pay high ransoms. Consequently, hackers might have less incentives to target such institutions to begin with. Due to these forces, hackers might target institutions at random. As a result, the equilibrium distribution of cyberattacks might closely resemble the size distribution of institutions within the economy.[3]

From an empirical perspective, scant data on both cyberattacks and private institutions poses a significant challenge. Besides the lack of public data on cyberattacks, it is difficult to find detailed information on private companies, many of which are

[1] The Global Risk Report of the World Economic Forum consistently reports cybersecurity threats among the top 10 concerns among world economic leaders. For more details, see https://www.weforum.org/publications/series/global-risks-report/. The Cybersecurity and Financial System Resilience Report of the Federal Reserve Board provides a descriptive account of policymakers' concerns about the potential system-wide repercussions of cyberattacks and measures taken to strengthen cybersecurity within the financial sector. For more details, see https://www.federalreserve.gov/publications/cybersecurity-and-financial-system-resilience-report.htm. Kashyap and Wetherilt (2019) emphasize the micro and macroprudential challenges posed by cyberattacks in modern economies.

[2] See Ablon (2018) for a description of hackers' motivations and their different types.

[3] See Dziubinski and Goyal (2013, 2017), Block, Dutta, and Dziubinski (2020) and Block, Chatterjee, and Dutta (2022) for equilibrium models of attack and defense within a network framework.

themselves victims of cyberattacks. To tackle this challenge, we combine a comprehensive data set on cyberattacks with the Walls & Associates (2020) National Establishment Time Series (NETS) data set to provide a representative description of the anatomy of cyberattacks across U.S. institutions.

With these data in hand—which account for about 2.5 million observations at the establishment-year level—we show that establishments of publicly traded companies are 2.68 times more likely to be targeted than establishments of nonpublic institutions—which, within our sample, cover non-publicly traded companies, non-profits, and government institutions. When compared with the average establishment in our sample, establishments generating 100 million dollars more in annual sales are 9.52% more likely to be targeted. And establishments with 100 more employees are 0.90% more likely to be victims of cyberattacks. Results at the institution level are even more striking. Publicly traded companies are 9.32 times more likely to be targeted than nonpublic institutions. And when compared with the average institution, institutions with 100 more employees are 2.12% more likely to be targeted. Our results control for various fixed-effects and are robust to variation in regression specifications and matching methodology.

Our findings are consistent with the idea that publicly traded companies and larger institutions are more likely targets of cyberattacks. Our analysis complements previous work examining cyber risk, including Jamilov, Rey, and Tahoun (2021), Kamiya, Kang, Kim, Milidonis, and Stulz (2021), and Florackis, Louca, Michaely, and Weber (2023). Although our paper and this literature share an emphasis on the distribution of cyberattacks, we provide a more granular picture of the anatomy of hacks across the whole distribution of U.S. establishments and not just publicly traded corporations. Our analysis also complements a literature examining the determinants of cyber risk—see, Aldasoro, Gambacorta, Giudici, and Leach (2020). Here, we provide a more detailed account of hacks across both public and private U.S. companies. Our results also complement a literature that emphasizes the potential system-wide implications of cyberattacks, including Duffie and Younger (2019), Kashyap and Wetherilt (2019), Kotidis and Schreft (2022), and Eisenbach, Kovner, and Lee (2022).

2 Data and Summary Statistics

To identify cyberattacks we use the Privacy Rights Clearinghouse (PRC) Data Breaches database—a collection of privacy breaches as reported by state Attorney Generals and the U.S. Department of Health and Human Services. Although these data contain over 9,000 observations spanning from 2005 through 2019, only a subsample of them affects U.S. institutions.[4] Because we are primarily interested in hacks—defined as breaches caused by an outside party or malware—affecting U.S. institutions, our initial sample contains 2,508 observations from 2005 to 2019. Besides institutions' names, observations in PRC provide the geographical location (city and state) of hacked institutions as well as the year of the hack.

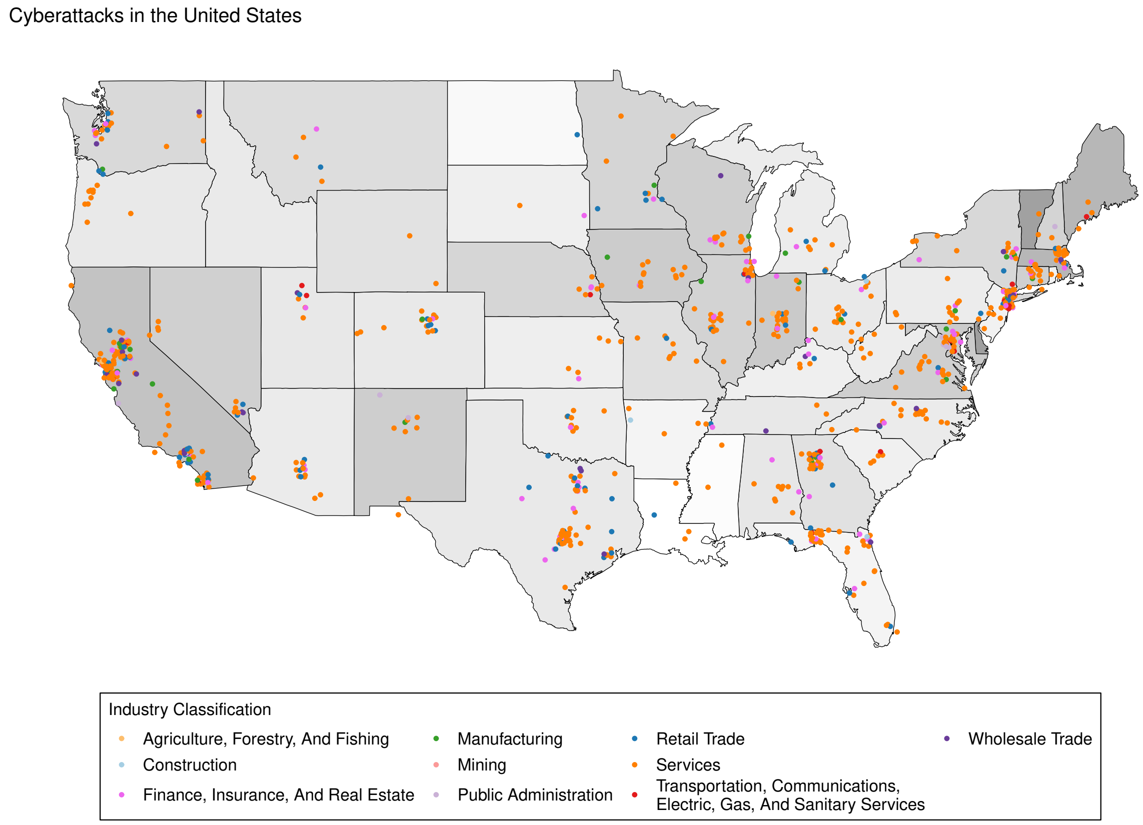
Because many observations in PRC refer to non-publicly traded companies and government institutions, we resort to NETS—a representative inventory of U.S. businesses with granular information for almost 80 million (private and public) establishments—to obtain characteristics of hacked institutions within our initial sample.[5] Our merging methodology matches observations in our initial sample with NETS establishments by name and location. We purposely generate our match at the establishment-year level to exploit potential variation across establishments within institutions. This mapping also helps us tackle concerns regarding over-representation of large publicly traded companies accounting for a multitude of establishments across the U.S.

Out of the 2,508 initial observations, our matching process generates an intermediate sample of 1,220 establishment-year observations for which we have detailed business information from 2005 to 2019.[6] Figure 1 depicts the geographical distribution of cyberattacks within our intermediate sample. Dots represent the location of hacked

[4] Because public information about cyber incidents is scant, the PRC data have been frequently used as a good approximation of cyber incidents in the U.S.; see Jamilov et al. (2021), Kamiya et al. (2021), and Floracks et al. (2023), among many others.

[5] Barnatchez, Crane, and Decker (2017) find that NETS can be a useful private-sector source of business microdata—relative to official U.S. business universe data sources—for studying business activity in granularity. Importantly, NETS can be accessed without extensive proposal, security clearance processes, and the need to be accessed inside of secure government facilities, potentially providing an efficient way to conduct research when business-level microdata is needed.

[6] Out of the 2508 observations in PRC, 19 observations lack any type of location data, 24 observations do not have state information, and 686 observations do not have city information. And 28% of observations have at least one piece of location data missing where only 1817 observations have name, city, and state information. In our baseline sample, we match 1396 observations in PRC. Hence, considering that only 1817 observations have name, city, and state information, our approximate matching rate is closer to 77%. A more detailed description of our merging methodology appears in Section A of the Online Appendix.


Cyberattacks in the United States
Industry Classification
● Agriculture, Forestry, And Fishing
● Construction
● Finance, Insurance, And Real Estate
● Manufacturing
● Mining
● Public Administration
● Retail Trade
● Services
● Transportation, Communications, Electric, Gas, And Sanitary Services
● Wholesale Trade

Note: Shading indicates proportion of attachs in each state, relative to total establishments.
A darker color indicates a greater proportion of establishments attacked. Each dot represents one cyberattack.

Figure 1: Distribution of hacks across the U.S. within our sample.

establishments. And colors are assigned according to their SIC division. States are colored according to the fraction of their establishments affected by cyberattacks in our sample—the lighter the color, the lower the fraction. Although many hacks in our sample affect establishments in California, Texas, New York, and Florida, Figure 1 shows that cyberattacks are somewhat equally spread across states. Figure 1 also shows that hacks affect establishments across a variety of different economic sectors.

Because observing more hacks affecting establishments with a specific characteristic might be just a reflection of the fact that there are more establishments with such characteristic, we combine the above data with a large random sample of NETS establishments. Our idea is to add controls and improve the representativeness of our data, obtaining a better picture of the anatomy of cyberattacks across U.S. establishments. In particular, we add a random sample of about 415,000 NETS establishments to our intermediate data. As a result, we obtain a sample with about 2.5 million establishment-year observations. For each establishment, we retrieve detailed information, including business location, headquarters, employment, sales, and other establishment-level data at the annual frequency, from 2005 to 2019.

Table 1 reports summary statistics of our baseline sample. Panel A reports statistics at the establishment-year level. The average establishment employs a bit less than seven employees per year and generates sales for about 690,000 dollars. On an average year, 2% of establishments belong to a publicly traded company. Panel B reports statistics at the institution-year level. As Panel B shows, the average institution in our sample employs a bit more than seven employees per year and generates sales of around 750,000 dollars. On an average year, 1% of institutions are publicly traded companies. The juxtaposition of Panels A and B shows that most institutions in our sample are small, non-publicly traded, and composed of, at most, one establishment.

Table 1:
Summary Statistics

This table reports statistics of establishments and institutions in our baseline sample at the annual frequency. Our sample contains 2,499,369 observations at the establishment-year level from 2005 to 2019. Panel A reports statistics at the establishment level while Panel B reports statistics at the institution level.

Panel A: Establishment level

	Mean	S.D.	10th	25th	50th	75th	90th
# of employees	6.89	13.92	1	2	2	5	15
Annual sales (in millions)	0.69	1.69	0.05	0.08	0.15	0.73	1.4
Ratio of public companies (in %)	2.4	0.31	2.1	2.2	2.3	2.5	2.8

Panel B: Institution level

	Mean	S.D.	10th	25th	50th	75th	90th
# of employees	7.21	16.57	1	2	2	4	13
Annual sales (in millions)	0.75	2.14	0.05	0.08	0.15	0.33	1.13
Ratio of public companies (in %)	1	0.13	0.8	0.92	0.99	1	1.2

3 Empirical Approach and Results

With our baseline sample in hand, we use the following logistic regression to explore whether an establishment's characteristics can alter the likelihood of being the target of a cyberattack:

$$\log\left(\frac{p_{it}}{1-p_{it}}\right) = \beta' X_{it} + \epsilon_{it}, \tag{1}$$

where there are observations on establishments (i) across years (t). The above equation uncovers a relationship between the logarithmic odds ratio, $\log\left(\frac{p_{it}}{1-p_{it}}\right)$, and establishment i's characteristics—wherein p_{it} captures the likelihood that establishment i is hacked at year t. Here, X_{it} is a vector of explanatory and control variables, which includes the constant term, and ϵ_{it} represents the error term. Explanatory variables include two measures of size—annual sales and number of employees—and whether an establishment belongs to a publicly traded company. To control for unobserved heterogeneity associated with characteristics at the industry-, state-, and year-levels, we include industry-, state-, and year-fixed effects. We also cluster standard errors at the industry-state level to correct for potential autocorrelation among residuals.

Table 2 reports our central findings. Panel A presents results at the establishment level while Panel B reports results at the institution level. For completeness, the first 6 columns in both Panels report different subsets of our explanatory variables while our most robust specifications are reported in columns 7. As Table 2 shows, our explanatory variables are statistically significant across most specifications. Panel A shows that larger establishments and establishments of publicly traded companies are more likely to be targeted. Panel B shows that this result also holds at the institution level.

As column 7 of Panel A shows, the coefficient associated with the public/private dummy (0.98748) is statistically significant. The same applies to the coefficients associated with sales (0.00091) and the number of employees (0.00009). Within a logistic regression framework, these values mean that establishments of publicly traded companies are 2.68 times more likely to be targeted than establishments of government or non-publicly traded institutions. When compared with the average establishment in our sample, establishments generating 100 million more in annual sales are 9.52% more likely to be targeted. And establishments with 100 more employees are 0.90% more likely to be victims of cyberattacks.

Largely consistent with Panel A's findings, Panel B shows that publicly traded companies and larger institutions are more likely to be targets. After converting coefficients to their exponentiated values for interpretation, Panel B shows that publicly traded companies are 9.32 times more likely to be targeted than nonpublic institutions. And, when compared to the average institution, institutions with 100 more employees are 2.12% more likely to be targeted.[7]

Table 2:
Central Findings

	Dependent variable: $\log(p_{it}/(1-p_{it}))$						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Results at the establishment level							
Public/Private dummy	1.02847***			0.98934***	1.01025***		0.98748***
	(0.31518)			(0.31257)	(0.31546)		(0.16251)
Sales		0.00132***		0.00114***		0.00107***	0.00091***
		(0.00021)		(0.00023)		(0.00021)	(0.00022)
# employees			0.00016***		0.00014***	0.00010***	0.00009***
			(0.00003)		(0.00003)	(0.00004)	(0.00002)
Observations	2,341,562	2,341,562	2,341,562	2,341,562	2,341,562	2,341,562	2,341,562
R-squared	0.08684	0.08479	0.08430	0.08766	0.08730	0.08485	0.08771
Panel B: Results at the institution level							
Public/Private dummy	2.31050***			2.19597***	2.22263***		2.23217***
	(0.33088)			(0.33184)	(0.32441)		(0.17962)
Sales		0.00083***		0.00066***		0.00016	-0.00008
		(0.00021)		(0.00018)		(0.00024)	(0.00014)
# employees			0.00021***		0.00020***	0.00019***	0.00021***
			(0.00002)		(0.00001)	(0.00002)	(0.00003)
Observations	2,242,172	2,242,172	2,242,172	2,242,172	2,242,172	2,242,172	2,242,172
R-squared	0.11154	0.10034	0.10435	0.11435	0.11903	0.10432	0.11894
Controls:							
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

3.1 Discussion and Data Limitations

Taken together, our findings support the view that larger establishments are more likely to become targets of cyberattacks. The same applies to establishments of publicly traded companies. Although many nonpublic and small institutions are frequent targets, not taking into consideration the overall distribution of establishments within the economy can lead to wrong conclusions.

Although PRC is, to the best of our knowledge, the most comprehensive public data on cyberattacks, we are mindful of its selection biases. Small and private companies could be underrepresented as they might not have the technology to uncover cyberattacks or may not bother to report them to authorities. Large and publicly traded companies might also have incentives to underreport—see, for example, Kamiya et al. (2021). In addition, larger and publicly traded companies might have the resources to hide these

[7] Section C in the Online Appendix shows that our findings are consistent with results from running probit (instead of logit) regressions and robust to variations in our merging methodology.

incidents. Because the selection bias can potentially go in either direction we do not take a stance on it and use the PRC data as it is.

4 Conclusion

Using granular data on cyberattacks and establishments in the United States we study whether and how an institution's characteristics can alter the likelihood of being the target of cyberattacks. We find that larger establishments—in terms of sales and number of employees—and establishments of publicly traded companies are more likely targets. A similar result holds at the institution level. Our results are robust to variation in regression specifications and merging methodologies.

References

Ablon, Lillian, 2018, Data thieves: The motivations of cyber threat actors and their use and monetization of stolen data, Testimony presented before the House Financial Services Committee, Subcommittee on Terrorism, and Illicit Finance, on March 15, 2018.

Aldasoro, Iñaki, Leonardo Gambacorta, Paolo Giudici, and Thomas Leach, 2020, The drivers of cyber risk, *BIS Working Paper* .

Barnatchez, Keith, Leland D. Crane, and Ryan A. Decker, 2017, An assessment of national establishment time series (nets) database, *Finance and Economics Discussion Series (FEDS)* .

Block, Francis, Kalyan Chatterjee, and Bhaskar Dutta, 2022, Attack and interception in networks, *Theoretical Economics* 18, 1511–1546.

Block, Francis, Bhaskar Dutta, and Marcin Dziubinski, 2020, A game of hide and seek in networks, *Journal of Economic Theory* 190.

Duffie, Darrell, and Joshua Younger, 2019, Cyber runs: How a cyber attack could affect u.s. financial institutions, *Hutchins Center Working Paper* 51.

Dziubinski, Marcin, and Sanjeev Goyal, 2013, Network design and defense, *Games and Economic Behavior* 79, 30–43.

Dziubinski, Marcin, and Sanjeev Goyal, 2017, How to defend a network?, *Theoretical Economics* 12, 331–376.

Eisenbach, Thomas M., Anna Kovner, and Michael Junho Lee, 2022, Cyber risk and the u.s. financial system: A pre-mortem analysis, *Journal of Financial Economics* 145, 802–826.

Florackis, Chris, Christodoulos Louca, Roni Michaely, and Michael Weber, 2023, Cybersecurity risk, *Review of Financial Studies* 36, 351–407.

Jamilov, Rustam, Hélene Rey, and Ahmed Tahoun, 2021, The anatomy of cyber risk, *NBER Working Paper Series* .

Kamiya, Shinichi, Jun-Koo Kang, Jungmin Kim, Andreas Milidonis, and Rene M. Stulz, 2021, Risk management, firm reputation, and the impact of successful cyberattacks on target firms, *Journal of Financial Economics* 139, 719–749.

Kashyap, Anil K., and Anne Wetherilt, 2019, Some principles for regulating cyber risk, *AEA Papers and Proceedings* 109, 482–487.

Kotidis, Antonis, and Stacey L. Schreft, 2022, Cyberattacks and financial stability: Evidence from a natural experiment, *Finance and Economics Discussion Series (FEDS)* .

Walls & Associates, 2020, National establishment time-series (nets) database.

Online Appendix for "On the Anatomy of Cyberattacks"

This Appendix contains material to supplement the analysis in "On the Anatomy of Cyberattacks." Section A provides more details about the construction of our baseline sample while Section B provides additional statistics. Section C shows that our central findings are robust to variation in regression specifications and merging methodologies.

A Construction of Baseline Sample

This section describes how to construct our baseline sample. We use the Privacy Rights Clearinghouse (PRC) Data Breaches database to uncover hacks in the U.S. Although these data contain 9,015 privacy breaches spanning from 2005 through 2019, only a subsample of those observations are hacks that affect U.S. institutions. After applying this filter, we obtain an initial sample of 2,508 observations at the establishment-year level from 2005 to 2019.

With these data in hand, we standardize names of institutions, cities, and states as several observations in PRC have misspelled names. This standardization also removes extra punctuation, numbers, and common words and helps us to match these data with NETS. Because observations in NETS are at the establishment level, we can find information at the precise location of the hack—which is especially important for large corporations with several establishments across the U.S. We first match observations in our initial sample with NETS establishments using their precise names. We corroborate that those matches are correct by comparing city and state information.

For observations with names that are slightly different from the names of NETS establishments, we use string-matching. Here, we use the Jaro-Winkler distance to assess the quality of our matches. Our baseline sample considers matches with a distance of less than or equal to 0.13. We purposely select this threshold as it helps to generate a sufficiently large number of precise matches. Section C explores the impact in our results of modifying this threshold. For those cases in which our algorithm associates a single establishment with multiple NETS establishments, we manually corroborate that our matches are correct. When these matches look sufficiently similar, we average NETS establishments' characteristics to create a hypothetical NETS establishment to which we associate the establishment from PRC. This process yields a sample of 26,410 potential matches, representing 1,220 different establishments.[1]

To generate a representative sample of the U.S. economy, we augment our intermediate sample with a large random draw of 413,139 different establishments from NETS.

[1] Although PRC contains 2508 hacks, many observations lack sufficient data to pin down the precise NETS establishment involved in a hack. Our most precise matching algorithm requires name, city, and state information to establish a mapping between an institution in PRC and a NETS establishment. When there is no city information, our algorithm still works, but it is less reliable as we need to filter establishments by state when querying information from NETS; here, the number of potential matches tends to be large. Hence, it is harder to differentiate between different matches when city data is missing. In our baseline sample, we match 1396 observations in PRC. Considering that only 1817 observations have name, city, and state information, our approximate matching rate is closer to 77% ($\approx$ 1396/1817). With these data in hand, we further require that establishments mentioned in PRC have sufficiently enough information on NETS so we can run panel regressions. Unfortunately, not all establishments mentioned in PRC satisfy that requirement. In particular, 176 hacks are matched to NETS establishments that lack sufficient information on employees, sales, industry, or business type. That is how we arrive to our final number of matched establishments: 1220.

This large random draw helps us tackle the dimensionality challenges inherent from using NETS.[2] For each establishment in these data, we retrieve the following annual information from 2005 to 2019: number of employees, sales, and whether an establishment is public or not. Within the baseline sample, we fill missing information with linear interpolation. Section C explores the impact in our results of modifying the way we fill in missing information. This process generates our baseline sample, which contains 2,499,369 observations at the establishment-year level from 2005 to 2019, representing 393,317 different establishments. Because certain establishments disappear throughout the sample, our baseline data set can be thought of an unbalanced panel.

B Additional Summary Statistics

This section provides additional statistics of our baseline sample. Figure 1 depicts the time series of hacks in our baseline sample separated by whether a hacked establishment is public or not, its number of employees, sales, and industry at the year of the hack. The figure on the upper left panel shows that most hacked establishments in our sample are either private or government institutions. The figure on the upper right panel shows that many hacked establishments have fewer than 20 employees when hacked. The figure on the lower left panel shows that several hacked establishments generate more than 3 million USD in annual sales. The figure on the lower right panel shows that a large fraction of hacked establishments is within the health, educational, and business sectors.

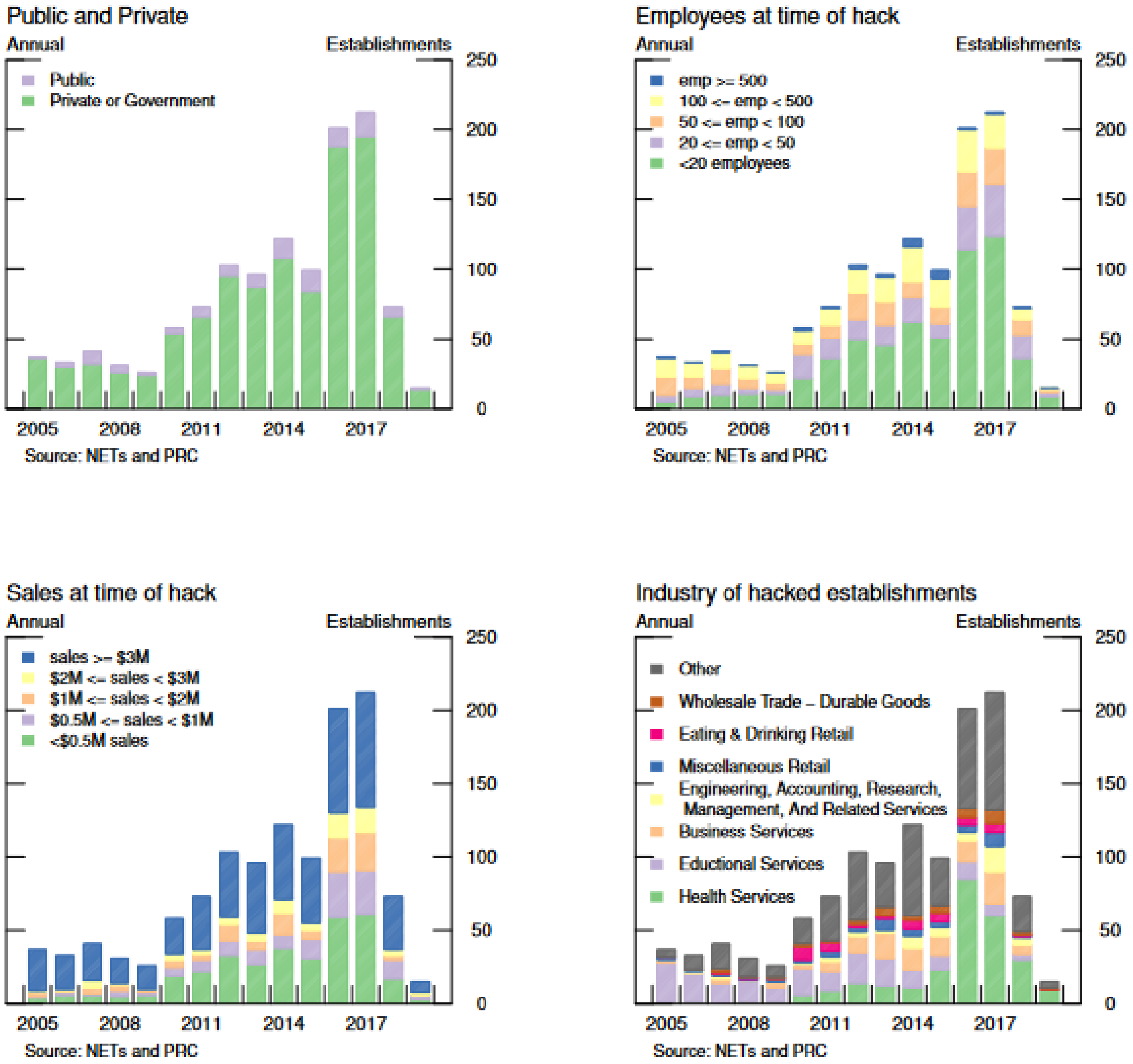
For completeness, Figure 2 provides other (potentially relevant) characteristics of the time series of hacked establishments in our baseline sample, including legal status, whether an establishment is an importer or exporter, whether an establishment has a government contract, and the gender of its CEO (or the CEO of its parent institution). Because most of these characteristics are less populated in NETS, we do not use them as controls within our regression specifications. Yet we present this information for the interested reader.

For the interested reader, Table 1 provides a more detailed breakdown of hacked establishments and institutions by major industry sector—which are captured by SIC divisions (i.e., 1-digit SIC codes).

C Robustness Tests

This section shows that our main results are robust to variation in regression specifications and merging methodologies. Section C.1 shows that our findings are consistent with results obtained from running probit regressions. Section C.2 shows that our results are robust to the way we deal with missing observations when creating our baseline sample. Section C.3 shows that our results are also robust to variation in the precise threshold used when string-matching data from PRC and NETS.

[2] NETS has 87,564,680 unique IDs. Each ID aims to represent a different establishment. If we were to load information for one year and all IDs would take about one hour and require 44.34 GB in memory. The dimensionality challenge arises because we run regressions using 15 years of data (2005-2019). To create our baseline sample, we pull data for each establishment from 4 different NETS data frames: Employees, Sales, SIC, and Misc. Using the complete NETS sample would require loading 1,313,347,200 observations—assuming each ID has 15 years of data—which is computationally unfeasible.



Public and Private

Source: NETs and PRC

Employees at time of hack

Source: NETs and PRC

Sales at time of hack

Source: NETs and PRC

Industry of hacked establishments

Source: NETs and PRC

Figure 1: Characteristics of hacked establishments.

C.1 Probit Specifications

Table 2 reports results from running probit regression using our baseline sample.

C.2 Dealing with Missing Information

Beside linear interpolation, we use two methodologies when dealing with missing information. Our first methodology simply omits observations with missing values. Table 3 reports these results.

Our second methodology uses linear interpolation for intermediate values while keeping extreme values constant when extrapolating two years ahead and backwards. For example, suppose the number of employees of a given establishment is known at years $(t - 1)$ and $(t + 1)$, but unknown at any other year. Using linear interpolation and the number of employees for years $(t - 1)$ and $(t + 1)$, we determine the number of employees for year t. We assume that values for years $(t-3)$ and $(t-2)$ equal the number of employees for year $(t - 1)$ and that values for years $(t + 2)$ and $(t + 3)$ equal the value

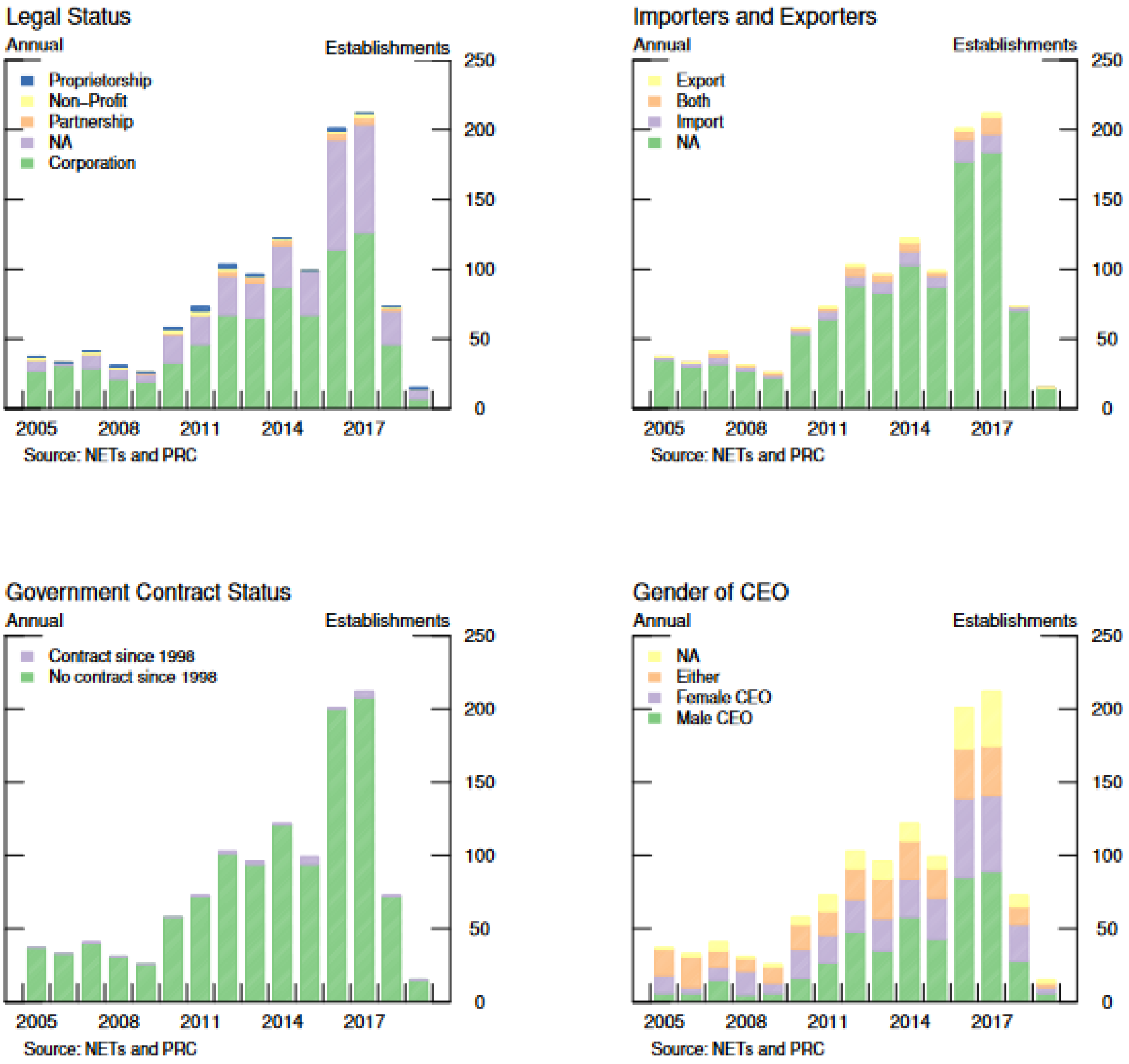


Figure 2: Other characteristics of hacked establishments in our sample.

for year $(t + 1)$. Table 4 reports these results.

For completeness, we also report results from running probit regressions using the above two methodologies. Table 5 reports the probit counterpart of Table 3 while Table 6 reports the probit counterpart of Table 4. As both Tables show, our central findings are robust to variation in how we handle missing observations.

C.3 Thresholds Used in String-Matching

The following tables show that our central findings are also robust to variations in the precise value used when string-matching observations from PRC and NETS. The baseline sample uses a Jaro-Winkler distance threshold of 0.13. Higher thresholds generate more potential matches, at the risk of being less precise. Lower thresholds generate more precise but fewer matches. We test two different thresholds. We use a threshold of 0.065 to increase the matching precision without compromising the size of our baseline sample too much. We also use a threshold of 0.26 to increase the size of our sample without

Table 1:
Industry Groups

This table reports statistics at the establishment and institution-level for observations in our the baseline sample.

Industry (SIC division)	# Establishments	# Institutions
Mining	596	545
Public administration	2803	1733
Manufacturing	11907	11558
Agriculture, forestry, and fishing	12351	11916
Wholesale trade	14770	14179
Transportation and public utilities	16782	15618
Construction	28768	28476
Finance, insurance, and real estate	36094	34515
Retail trade	49331	44955
Services	219983	210915

Table 2:
Probit regressions using baseline sample

	Probit Regressions						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Panel A: Without fixed-effects						
Public/Private dummy	0.43320***			0.41509***	0.42445***		0.41329***
	(0.03050)			(0.03098)	(0.03069)		(0.03102)
Sales		0.00059***		0.00048***		0.00048**	0.00038***
		(0.00006)		(0.00006)		(0.00007)	(0.00007)
# employees			0.00008***		0.00007***	0.00005***	0.00004***
			(0.00001)		(0.00001)	(0.00001)	(0.00001)
Observations	2,499,369	2,499,369	2,499,369	2,499,369	2,499,369	2,499,369	2,499,369
	Panel B: With fixed-effects & clustered standard errors						
Public/Private dummy	0.3346***			0.31944***	0.32689***		0.31893***
	(0.09916)			(0.09747)	(0.09887)		(0.09746)
Sales		0.00058***		0.00051***		0.00047***	0.00041***
		(0.00010)		(0.00009)		(0.00010)	(0.00010)
# employees			0.00007***		0.00006***	0.00004**	0.00003**
			(0.00002)		(0.00001)	(0.00002)	(0.00003)
Observations	2,341,562	2,341,562	2,341,562	2,341,562	2,341,562	2,341,562	2,341,562
R-squared	0.08808	0.08609	0.08542	0.08923	0.08873	0.08614	0.08927

Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

compromising the precision of our matching too much.

C.3.1 Threshold 0.065

Table 7 reports results when 0.065 is used as a distance threshold when performing string-matching.

C.3.2 Threshold 0.26

Table 8 reports results when 0.26 is used as a distance threshold when performing string-matching.

Table 3:
Results when missing values are omitted. Logit regressions.

	Dependent variable: $\log(p_{it}/(1-p_{it}))$						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Panel A: Without fixed-effects						
Public/Private dummy	1.62464***			1.57539***	1.60508***		1.56742***
	(0.09566)			(0.09740)	(0.09632)		(0.09768)
Sales		0.00141***		0.00105***		0.00124**	0.00091***
		(0.00013)		(0.00013)		(0.00014)	(0.00015)
# employees			0.00015***		0.00012***	0.00010***	0.00008***
			(0.00002)		(0.00002)	(0.00002)	(0.00002)
Observations	2,510,338	2,510,338	2,510,338	2,510,338	2,510,338	2,510,338	2,510,338
	Panel B: With fixed-effects & clustered standard errors						
Public/Private dummy	1.25150***			1.21458***	1.23144***		1.21128***
	(0.32018)			(0.31935)	(0.32069)		(0.16577)
Sales		0.00134***		0.00116***		0.00108***	0.00093***
		(0.00022)		(0.00023)		(0.00022)	(0.00023)
# employees			0.00017***		0.00014***	0.00010**	0.00009***
			(0.00003)		(0.00003)	(0.00004)	(0.00003)
Observations	2,348,581	2,348,581	2,348,581	2,348,581	2,348,581	2,348,581	2,348,581
R-squared	0.08953	0.08592	0.08522	0.09070	0.09017	0.08602	0.09078

Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

Table 4:
Results when missing values are interpolated. Logit regressions.

	Dependent variable: $\log(p_{it}/(1-p_{it}))$						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Panel A: Without fixed-effects						
Public/Private dummy	1.50593***			1.46282***	1.48714***		1.45264***
	(0.09809)			(0.09964)	(0.09879)		(0.10005)
Sales		0.00149***		0.00117***		0.00135***	0.00105***
		(0.00014)		(0.00015)		(0.00015)	(0.00016)
# employees			0.00015***		0.00012***	0.00010***	0.00008***
			(0.00002)		(0.00002)	(0.00002)	(0.00002)
Observations	2,370,419	2,370,419	2,370,419	2,370,419	2,370,419	2,370,419	2,370,419
	Panel B: With fixed-effects & clustered standard errors						
Public/Private dummy	1.13705***			1.11053***	1.11234***		1.09918***
	(0.34388)			(0.34284)	(0.34685)		(0.20103)
Sales		0.00210***		0.00204***		0.00178**	0.00177**
		(0.00072)		(0.00069)		(0.00074)	(0.00073)
# employees			0.00018***		0.00015***	0.00011***	0.00009***
			(0.00004)		(0.00004)	(0.00003)	(0.00003)
Observations	2,212,604	2,212,604	2,212,604	2,212,604	2,212,604	2,212,604	2,212,604
R-squared	0.10766	0.10595	0.10485	0.10956	0.10842	0.10617	0.10968

Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

Table 5:
Results when missing values are omitted. Probit regressions.

	Probit Regressions						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Panel A: Without fixed-effects						
Public/Private dummy	0.47953***			0.45973***	0.47093***		0.45770***
	(0.02981)			(0.03036)	(0.03002)		(0.03040)
Sales		0.00064***		0.00052***		0.00053***	0.00042***
		(0.00006)		(0.00006)		(0.00006)	(0.00007)
# employees			0.00008***		0.00007***	0.00005***	0.00004***
			(0.00001)		(0.00001)	(0.00001)	(0.00001)
Observations	2,510,338	2,510,338	2,510,338	2,510,338	2,510,338	2,510,338	2,510,338
	Panel B: With fixed-effects & clustered standard errors						
Public/Private dummy	0.40365***			0.38938***	0.39609***		0.38843***
	(0.10084)			(0.09943)	(0.10063)		(0.09938)
Sales		0.00061***		0.00054***		0.00050***	0.00044***
		(0.00010)		(0.00009)		(0.00010)	(0.00010)
# employees			0.00008***		0.00007***	0.00004**	0.00003**
			(0.00002)		(0.00002)	(0.00002)	(0.00002)
Observations	2,348,581	2,348,581	2,348,581	2,348,581	2,348,581	2,348,581	2,348,581
R-squared	0.09078	0.08726	0.08630	0.09244	0.09170	0.08735	0.09251

Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

Table 6:
Results when missing values are interpolated. Probit regressions.

	Probit Regressions						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Panel A: Without fixed-effects						
Public/Private dummy	0.44792***			0.42906***	0.43957***		0.42529***
	(0.03075)			(0.03126)	(0.03099)		(0.03137)
Sales		0.00077***		0.00066***		0.00067***	0.00057***
		(0.00008)		(0.00008)		(0.00008)	(0.00008)
# employees			0.00009***		0.00008***	0.00006***	0.00005***
			(0.00001)		(0.00001)	(0.00001)	(0.00001)
Observations	2,370,419	2,370,419	2,370,419	2,370,419	2,370,419	2,370,419	2,370,419
	Panel B: With fixed-effects & clustered standard errors						
Public/Private dummy	0.36782***			0.34999***	0.35864***		0.34695***
	(0.10878)			(0.10797)	(0.10944)		(0.10862)
Sales		0.00123***		0.00117***		0.00109***	0.00105***
		(0.00040)		(0.00036)		(0.00042)	(0.00037)
# employees			0.00009**		0.00008**	0.00004	0.00004
			(0.00004)		(0.00004)	(0.00003)	(0.00002)
Observations	2,212,604	2,212,604	2,212,604	2,212,604	2,212,604	2,212,604	2,212,604
R-squared	0.11000	0.10953	0.10726	0.11323	0.11114	0.10973	0.11335

Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

Table 7:
Logit regressions. Threshold: 0.065.

	Dependent variable: log($p_{it}/(1-p_{it})$)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Panel A: Without fixed-effects						
Public/Private dummy	1.52269***			1.48176***	1.50625***		1.47486***
	(0.10486)			(0.10641)	(0.10548)		(0.10667)
Sales		0.00139***		0.00105***		0.00123**	0.00093***
		(0.00015)		(0.00016)		(0.00017)	(0.00017)
# employees			0.00014***		0.00011***	0.00009***	0.00008***
			(0.00002)		(0.00002)	(0.00003)	(0.00003)
Observations	2,473,770	2,473,770	2,473,770	2,473,770	2,473,770	2,473,770	2,473,770
	Panel B: With fixed-effects & clustered standard errors						
Public/Private dummy	1.06187***			1.02741***	1.04545***		1.02659***
	(0.37061)			(0.36508)	(0.37134)		(0.15999)
Sales		0.00153***		0.00140***		0.00132***	0.00120***
		(0.00028)		(0.00024)		(0.00026)	(0.00021)
# employees			0.00016***		0.00014***	0.00007*	0.00007***
			(0.00003)		(0.00003)	(0.00004)	(0.00002)
Observations	2,308,832	2,308,832	2,308,832	2,308,832	2,308,832	2,308,832	2,308,832
R-squared	0.08475	0.08266	0.08197	0.08574	0.08515	0.08262	0.08570

Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

Table 8:
Logit regressions. Threshold: 0.26.

	Dependent variable: log($p_{it}/(1-p_{it})$)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Panel A: Without fixed-effects						
Public/Private dummy	1.27134***			1.23438***	1.25470***		1.22910***
	(0.07899)			(0.08004)	(0.07941)		(0.08016)
Sales		0.00115***		0.00087***		0.00096**	0.00071***
		(0.00012)		(0.00013)		(0.00014)	(0.00015)
# employees			0.00014***		0.00012***	0.00010***	0.00008***
			(0.00002)		(0.00002)	(0.00002)	(0.00002)
Observations	2,795,110	2,795,110	2,795,110	2,795,110	2,795,110	2,795,110	2,795,110
	Panel B: With fixed-effects & clustered standard errors						
Public/Private dummy	0.97248***			0.94040***	0.95336***		0.93721***
	(0.29309)			(0.29280)	(0.29344)		(0.16944)
Sales		0.00091***		0.00071***		0.00063***	0.00045**
		(0.00020)		(0.00018)		(0.00020)	(0.00021)
# employees			0.00016***		0.00014***	0.00012***	0.00011***
			(0.00003)		(0.00003)	(0.00003)	(0.00002)
Observations	2,677,613	2,677,613	2,677,613	2,677,613	2,677,613	2,677,613	2,677,613
R-squared	0.06387	0.06156	0.06148	0.06422	0.06424	0.06168	0.06432

Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.