On the Anatomy of Cyberattacks*

Jin-Wook Chang Kartik Jayachandran Carlos A. Ramírez

Ali Tintera

February 27, 2024

Abstract

Using detailed information on cyberattacks and establishments in the United States, we study how an establishment's characteristics can alter the likelihood of being the target 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.

*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 constructing a new data set with detailed information on cyberattacks and establishments in the U.S. and studying which types of establishments are more likely to be targeted.

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 both hackers' motivation and the response of institutions' response, both of which are difficult to observe; see Ablon (2018) for a descriptive account of the multiple motivations of hackers and their different types. To understand this observation more clearly, consider a simple model economy wherein hackers are purely financially motivated. To increase their expected profits, assume hackers target larger institutions as they might obtain higher ransom payments from successful attacks. Taking into consideration this strategy, larger institutions would increase their investments in cybersecurity, making it more difficult for hackers to implement successful attacks. Therefore, targeting such institutions may become less profitable (as successful attacks become less likely). A similar idea applies to smaller institutions. Here, however, expected profits from successful attacks might be smaller, as smaller institutions might not be able to pay ransoms as high as larger institutions. Consequently, hackers might have less incentives to target such institutions to begin with. Due to these forces, hackers might decide to target institutions at random in equilibrium. Consequently, the distribution of cyberattacks might closely resemble the size distribution across institutions within the economy. See Dziubinski and Goyal (2013, 2017) and Block, Dutta, and Dziubinski (2020); Block, Chatterjee, and Dutta (2022) for equilibrium models of attack and defense

[1]The Global Risk Perception Survey consistently reports cybersecurity threats among the top 10 concerns among world economic leaders. For more details, see World Economic Forum (2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023). In addition, see Board of Governors of the Federal Reserve System (2021, 2022, 2023) for a descriptive account of policymakers' concerns about the potential system-wide repercussions of cyberattacks and measures taken to strengthen cybersecurity within the financial sector. Kashyap and Wetherilt (2019) emphasize the micro and macroprudential challenges posed by cyberattacks in modern economies.

within a network framework.

From an empirical perspective, scant data on both cyberattacks and private institutions poses a significant challenge. In addition to lack of public data on cyberattacks, it is difficult to find detailed information on private institutions, many of which are themselves victims of cyberattacks. To tackle this challenge, we construct a new data set that combines a somewhat comprehensive data set on cyberattacks that is publicly available with the National Establishment Time Series (NETS) data set—which contains granular information on business activities for a large fraction of U.S. (private and public) establishments—to uncover 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 public institutions are 2.68 times more likely to be targeted than establishments of nonpublic 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 institutional level are even more striking. Public institutions 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 are robust to a multitude of controls as well as variation in our regression specification and matching methodology.

Our findings are consistent with the idea that public and larger institutions are more susceptible to cyberattacks. Within an interconnected economy, cyberattacks may therefore pose system-wide risk. Our analysis complements previous work examining cyber risk, including Jamilov, Rey, and Tahoun (2021), Kamiya, Kang, Kim, Milidonis, and Stulz (2021), Florackis, Louca, Michaely, and Weber (2023). Although our paper and this literature share an emphasis on the distribution of cyberattacks, we seek to provide a more granular picture of the anatomy of hacks across the whole distribution of U.S. establishments and not just public corporations. Our analysis also complements a literature examining the determinants of cyber risk. Compared with Aldasoro, Gambacorta, Giudici, and Leach (2020), we provide a more detailed account of hacks across both public and private U.S. establishments. 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.[2] 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 private institutions, we resort to NETS—a representative inventory of U.S. businesses with granular information for almost 80 million (private and public) U.S. establishments—to obtain characteristics of hacked institutions within our initial sample.[3] 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 the same institution. This mapping also helps us tackle concerns regarding over-representation of large institutions 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. Figure 1 depicts the geographical distribution of

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

[3]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.

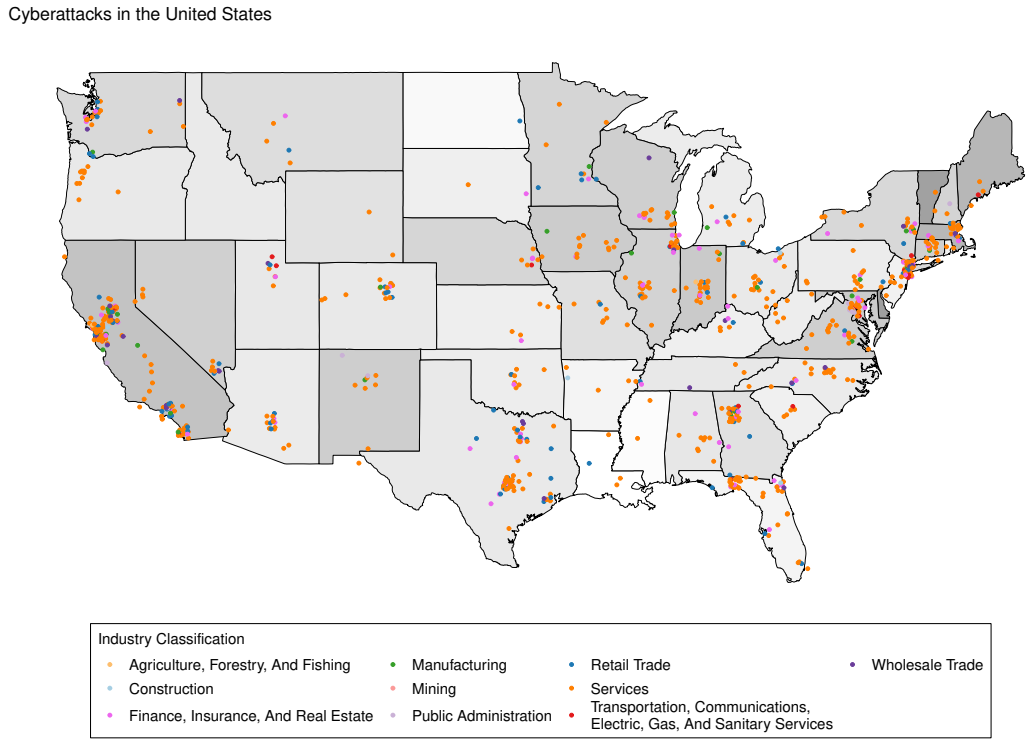

Figure 1: Distribution of hacks across the U.S. within our sample.

cyberattacks within our intermediate sample. Dots represent the location of hacked establishments. And their colors are assigned according to the SIC division of hacked establishments. To illustrate the population of establishments within states, 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 in our sample. 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 in the economy 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 locations, 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. The 25th and 75th percentiles of the annual distribution of employees are two and five workers, respectively. And the 25th and 75th percentiles of the annual distribution of sales are 80,000 and 730,000 dollars, respectively. On an average year, 2% of establishments belong to a publicly traded institution. Panel B reports statistics at the institution-year level. This exercise is potentially revealing as certain institutions—especially large corporations—might be represented by a multitude of different establishments spread across states. 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. The juxtaposition of Panels A and B shows that most institutions in our sample are small, nonpublic, 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 allows us to uncover a relationship between the logarithmic odds ratio, $\log\left(\frac{p_{it}}{1-p_{it}}\right)$, and establishment i' 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 include 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 public institution. To correct for unobserved heterogeneity associated with characteristics at the industry-, state-, and year-levels, we include industry-, state-, and year-fixed effects in our baseline specifications. In such specifications, 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 without fixed effects while Panel B reports our baseline specifications that include fixed effects and clustered standard errors. For completeness, the first 6 columns in both panels report different subsets of our explanatory variables while our most robust specifications are reported in column 7. As Table 2 shows, our explanatory variables are statistically significant across all specifications. Panel A shows that larger establishments and establishments of public institutions are more likely to be targeted. Panel B shows that this result holds even after correcting for state-, industry-, and year-fixed effects, and clustering standard errors. As column 7 of Panel B shows, establishments of public institutions are 2.68 times more likely to be targeted than establishments of government or private 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.

Table 3 reports results at the institutional level when variation across establishments within the same institution is absent. Largely consistent with the previous findings,

Table 2: Results at establishment level

	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	1.46168***			1.41856***	1.44329***		1.4114***
	(0.09807)			(0.09965)	(0.09871)		(0.09991)
Sales		0.00132***		0.00099***		0.00115***	0.00085***
		(0.00014)		(0.00014)		(0.00016)	(0.00016)
# employees			0.00015***		0.00012***	0.00010***	0.00008***
			(0.00002)		(0.00002)	(0.00003)	(0.00003)
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	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

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

Table 3 shows that public and larger institutions are more likely to be targets. As before, Panel A presents results without fixed effects while Panel B reports specifications with fixed effects and clustered standard errors. We find that public institutions 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.

Table 3: Results at institution level

	Dependent variable: $\log(p_{it}/(1-p_{it}))$						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Without fixed-effects							
Public/Private dummy	2.67371***			2.47081***	2.48903***		2.53972***
	(0.09547)			(0.10264)	(0.09670)		(0.09885)
Sales		0.00113***		0.00073***		0.00031**	-0.00024*
		(0.00006)		(0.00007)		(0.00013)	(0.00012)
# employees			0.00029***		0.00028***	0.00026***	0.00031***
			(0.00001)		(0.00001)	(0.00002)	(0.00002)
Observations	2,368,559	2,368,559	2,368,559	2,368,559	2,368,559	2,368,559	2,368,559
Panel B: With fixed-effects & clustered standard errors							
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

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 public institutions. 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.

While PRC is, to the best of our knowledge, the most comprehensive public data on cyberattacks, we are mindful of the limitations associated with using these data. One important limitation relates to selection biases. Small and private institutions could be underrepresented as they might not have the technology to uncover sophisticated cyberattacks or may not bother to report these incidents to authorities. Large and public institutions might also have incentives to underreport. While larger and public institutions are more likely to have the technology to uncover an incident, they are also more likely to be concerned about their reputation loss, headline risk, or stock price fluctuations—see, for example, Kamiya et al. (2021). In addition, larger and public institutions might have the resources to hide these incidents and address them by themselves. Consequently, although the selection bias is present, it can potentially go in either direction, so we take an agnostic stance on it and use the PRC data as it is.

4 Conclusion

We empirically study whether an institution's characteristics can alter the likelihood of being the target of cyberattacks. To do so, we construct a detailed data on cyberattacks and establishments in the United States. We find that larger establishments—in terms of sales and number of employees—and establishments of public institutions are more likely targets. A similar result holds at the institutional level. Our results are robust to a battery of controls as well as variation in regression specifications and merging methodologies. Largely aligned with the findings of Kotidis and Schreft (2022), Eisenbach et al. (2022), and Board of Governors of the Federal Reserve System (2021, 2022, 2023), our results are consistent with the idea that cyberattacks could impose significant adverse effects in modern interconnected economies wherein large public institutions might play a more significant role.

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.

Board of Governors of the Federal Reserve System, 2021, Cybersecurity and financial system resilience report.

Board of Governors of the Federal Reserve System, 2022, Cybersecurity and financial system resilience report.

Board of Governors of the Federal Reserve System, 2023, Cybersecurity and financial system resilience report.

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)* .

World Economic Forum, 2016, *The Global Risks Report*, 11th edition.

World Economic Forum, 2017, *The Global Risks Report*, 12th edition.

World Economic Forum, 2018, *The Global Risks Report*, 13th edition.

World Economic Forum, 2019, *The Global Risks Report*, 14th edition.

World Economic Forum, 2020, *The Global Risks Report*, 15th edition.

World Economic Forum, 2021, *The Global Risks Report*, 16th edition.

World Economic Forum, 2022, *The Global Risks Report*, 17th edition.

World Economic Forum, 2023, *The Global Risks Report*, 18th edition.

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 we 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 and 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 and 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.

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. This large random draw helps us tackle the dimensionality challenges inherent from using NETS. 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 2 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.

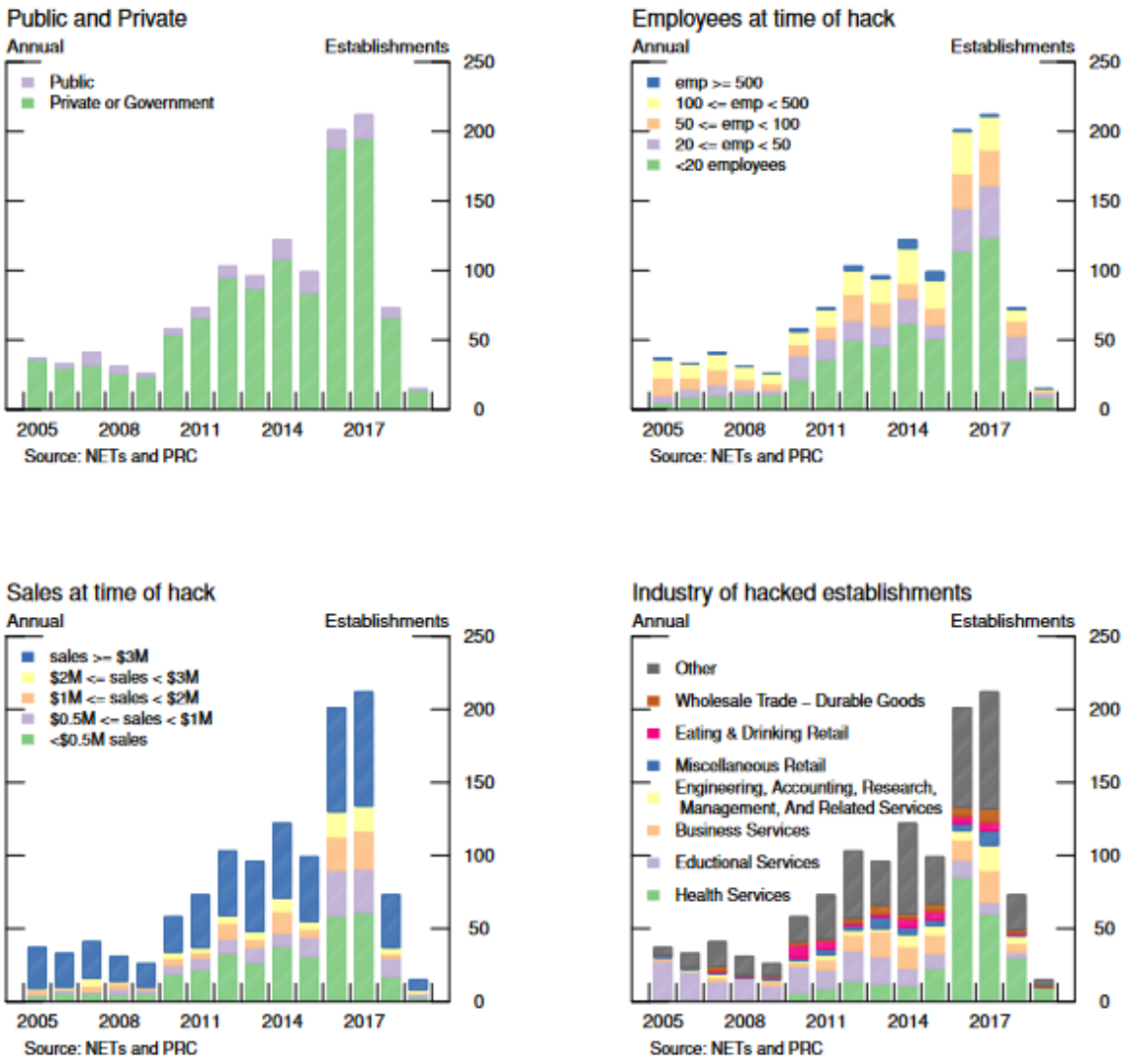

Figure 2: Characteristics of hacked establishments.

For completeness, Figure 3 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 here for the interested reader.

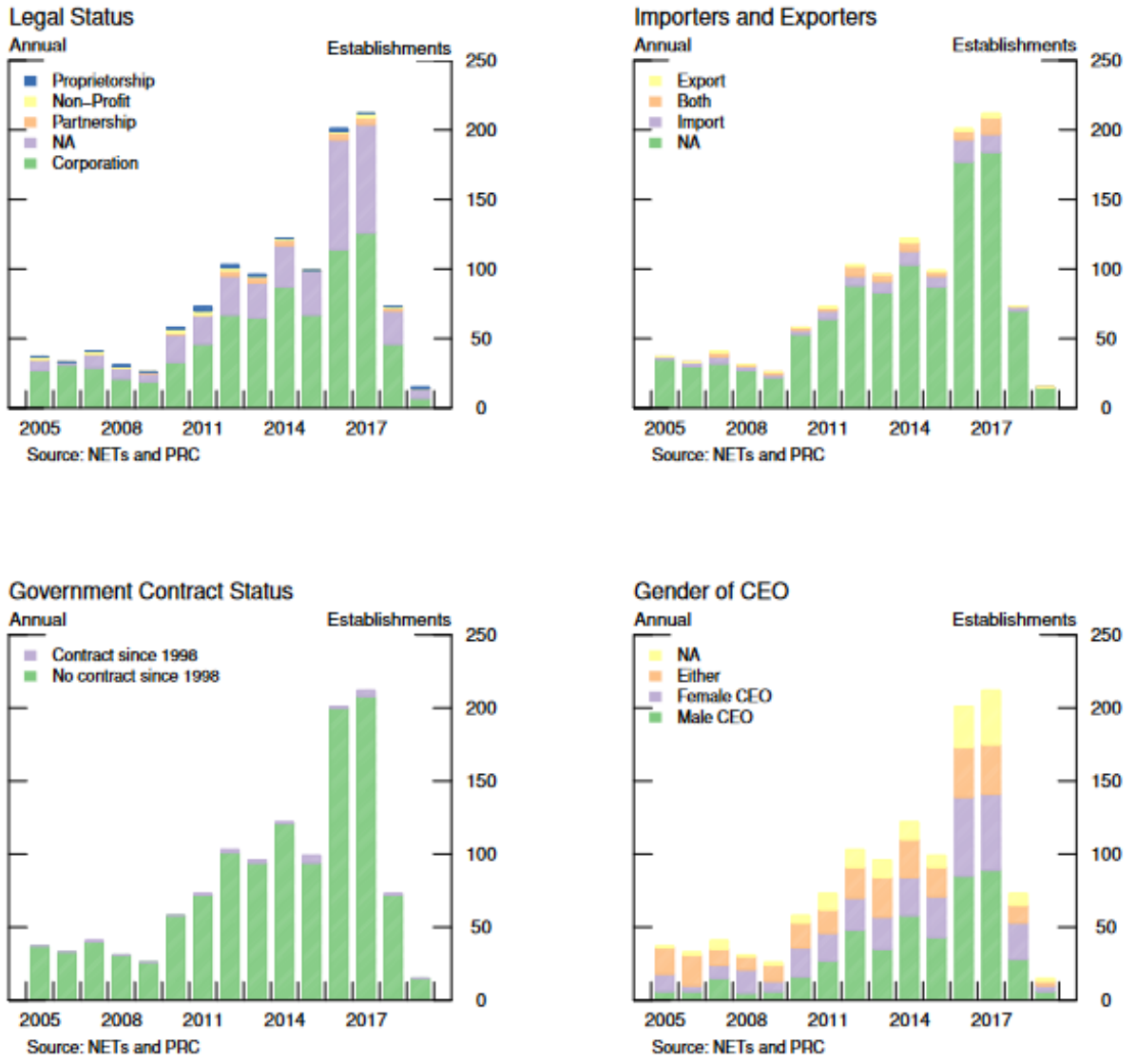

Figure 3: Other characteristics of hacked establishments in our sample.

For completeness, Figure 4 depicts the geographical distribution of all the hacks affecting U.S. establishments within the complete PRC data. Here, state colors are based on the number of hacks per state.

For the interested reader, Table 4 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

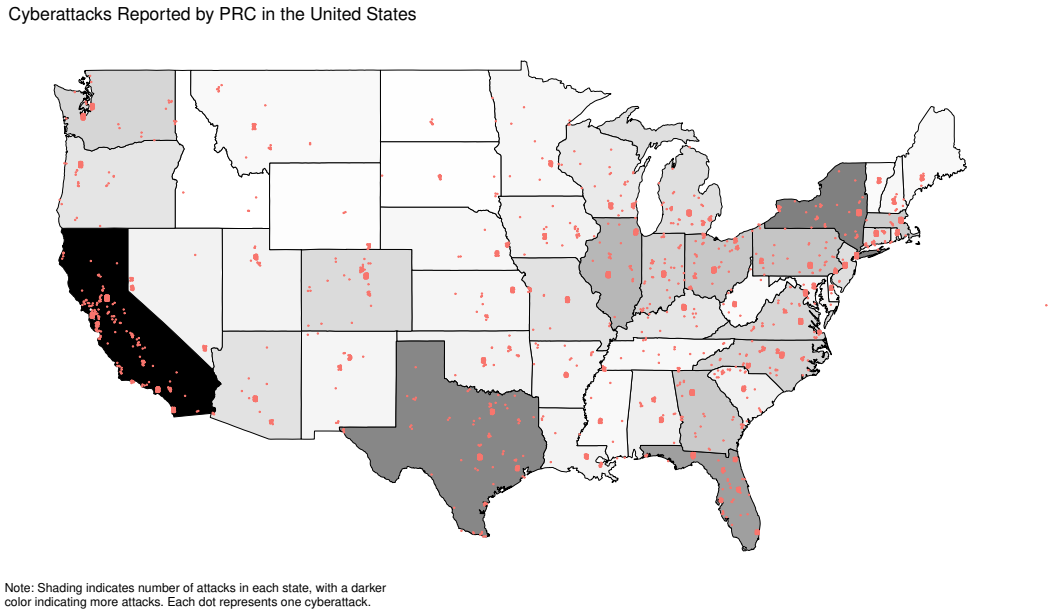

Figure 4: Distribution of hacks across the U.S. within PRC sample.

Table 4:
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

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 the precise threshold we use when string-matching data from PRC and NETS.

C.1 Probit Specifications

Table 5 reports results from running probit regression using our baseline sample.

Table 5: 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 * $p < 0.1$.

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 6 reports these results.

Table 6: 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 * $p < 0.1$.

Our second methodology uses linear interpolation for intermediate values while keeping 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 for year $(t+1)$. Table 7 reports these results.

Table 7: 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 * $p < 0.1$.

For completeness, we also report results from running probit regressions using the above two methodologies. Table 8 reports the probit counterpart of Table 6 while Table 9 reports the probit counterpart of Table 7. As both Tables show, our central findings are robust to variation in how we handle missing observations.

Table 8: 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 9: 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 * $p < 0.1$.

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 compromising the precision of our matching too much.

C.3.1 Threshold 0.065

Table 10 reports results when 0.065 is used as a distance threshold when performing string-matching.

Table 10: 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 * $p < 0.1$.

C.3.2 Threshold 0.26

Table 11 reports results when 0.26 is used as a distance threshold when performing string-matching.

Table 11: 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 * $p < 0.1$.