Crime in the Digital Age: Do Cyber Attacks Lead to Identity Theft? *

Claudio Mezzetti, Keshini Muthukuda, Haishan Yuan †

April 24, 2024

Abstract

We study whether data hacking of local organizations causes an increase in identity theft in the organization's local area. We use a difference in differences approach exploiting the timing of incidents of hacking and identity theft in the USA from 2015 to 2018, and estimate a fixed effects model that includes time and Core Based Statistical Area, or County, fixed effects. We find that a hacking incident in a local organization leads to 0.792 to 1.044 more identity thefts per 10,000 population in the local area the following year. The increase represents a 42% to 77% increase in the average prevalence of identity theft. We also show that among all our controls the unemployment rate is the most significant predictor of identity theft.

Keywords: Crime, Cybersecurity, Hacking Attacks, Identity Theft.

*For their valuable comments and suggestions, we would like to thank Sarah Bana, Andrea La Nauze, Lizy Yu and the audiences at the New Zealand Association of Economists (NZAE 2023) and the Australian Conference of Economists (ACE) 2023). Mezzetti's work was funded in part by Australian Research Council grant DP190102904.

†School of Economics, University of Queensland, Colin Clark Building 39, St. Lucia, 4072, QLD, Australia. Email: c.mezzetti@uq.edu.au; k.muthukuda@uq.net.au; h.yuan@uq.edu.au.

1 Introduction

As more economic activities have shifted online, corporations, non-profit organizations, and government agencies have amassed a vast amount of personal information. At the same time security breaches in their databases have become more and more common. They have been facilitated by inadequate legal deterrence of cybercriminals, difficulties in the attribution of legal responsibilities and the allocation of remedies, and underinvestment in cybersecurity.

Cybercriminals often operate in different regions from their victims (e.g., see Kshetri, 2010), because geographic proximity between the perpetrator and the victim is not necessary, unlike for traditional crimes, and the opportunity cost of committing a cybercrime is lower in low-income regions, whereas the return is higher if the victims are in high-income regions.[1] Thus, the legal deterrence of cybercrime and its policing require cooperation among law enforcement agencies from different regions. However, such collaboration, and especially international cooperation, remains a challenge. Indeed, it is a common view that some countries may tolerate or even sponsor cybercrime (see, e.g., Almond et al., 2020).

As many companies have accumulated large amounts of personal information that is stored online, the attribution of legal responsibilities for stolen data is difficult, as it is often legally unclear whom to blame. The correct allocation of remedies is also a challenging problem, because it is uncertain how to compensate the victims; different people have different valuations of privacy (Prince and Wallsten, 2022), and the causal link between specific data breaches and the financial damages from the data breaches may be difficult for even an attentive consumer to discern.

Companies tend to underinvest in cybersecurity, due to the positive externalities of a company's investment in cybersecurity on other companies (Anderson and Moore, 2006). Positive externalities arise because, by reducing the chances of a virus infection, increased cybersecurity in one company reduces the network contagion effects on other companies. They also arise because of the complementarity of personal information for the purpose of cybercrimes like identity theft. While a small piece of personal information may not be enough, a combination of pieces from different sources may be sufficient for identity theft. By protecting the piece of

[1] Low- and middle-income regions often lack: (*i*) the technical capacity to police cyberspace; (*ii*) good job prospects for technology savvy and educated individuals (Moore et al., 2009); (*iii*) the incentives to police cybercrimes that benefit local perpetrators and harm victims in other regions.

information that a company has with enhanced cybersecurity, the complementary pieces of information from other companies are also partially protected, as they become less valuable to hackers, thus decreasing the likelihood of attacks. Underinvestment in cybersecurity is also facilitated by the presence of users who fail to factor in the risks and potential costs of inadequate cybersecurity due to rational inattention, neglect, low probability of financial damages, lack of knowledge, or lack of alternatives in their service providers, such as educational and medical institutions. Only after cyberbreaches and the bad public relations damage that comes with them, do firms catch up on investing in cybersecurity (Bana et al., 2022). To make matters worse, managers typically delay and under-report data breaches (Amir et al., 2018).

Stolen information can be used by cyber criminals in different ways; it can, for example, be sold in the dark-web market. Even if illegally acquired, it could conceivably be used to the benefit of the affected, e.g., if it is used for efficiency enhancing, personalised, advertising. It is thus unclear, and natural to ask, how widespread the economic and emotional damages are for those whose personal information has been stolen by cyber criminals. Identity theft clearly imposes significant personal costs and measuring the impact of a hacking attack on identity theft, as we do in this paper, is a way to measure the economic and emotional damage on individuals of security breaches.

The Federal Bureau of Investigation (FBI) defines Identity Theft as "wrongfully obtaining and using another person's personal data (e.g., name, date of birth, social security numbers, driver's license numbers, etc.)."[2] The prevalence of digital payment systems linked to online identity has made identity theft increasingly attractive to criminals (Anderson et al., 2008; Roberds and Schreft, 2009; Kahn and Roberds, 2008; Mikhed and Vogan, 2018). According to the consumer sentinel network data book of the Federal Trade Commission (2021), there were 1,387,615 identity thefts reported in the U.S. in 2020, which represented an approximately 5-fold increase within the last ten years. A study by the Javelin Strategy & Research institution estimated the total identity fraud losses in 2020 to be up to $56 billion (Buzzard and Kitten, 2021).

Furthermore, the Consumer Plus study of TransUnion (2021) showed that during the fourth quarter of 2021, 42% of U.S consumers reported being targets of dig-

[2]U.S Department of Justice (2018, p.30)

ital frauds, out of which 22% was related to identity theft. Because identity thefts are often linked or perpetrated in conjunction with other frauds, such as financial frauds (opening up or using bank/credit accounts), welfare fraud, and tax frauds, the social cost of identity theft is likely to be significantly high.

Motivated by this development, in this paper, we ask whether data breaches (hacking) in an organization lead to a higher rate of identity theft in the area where the organization is located. We exploit the timing of successful hacking incidents that lead to the breaches of a large number of personal information that is likely to be disproportionately local.

Our empirical specification focuses on two geographic units, county and core-based statistical areas (CBSA). The latter include metropolitan and micropolitan statistical areas; both are urbanized areas, typically comprising of a central county and a couple of other counties that have strong ties with, and are geographically located near, the central county. On the one hand, because the degree of urbanization is a significant social predictor of crime in general, one might expect the impact of hacking on identity theft to be greater in CBSAs. On the other hand, CBSA are larger geographical areas, with greater average distance from the hacked organization and, since we focus on localized hacking attacks, one might expect the impact of hacking on identity theft to be greater in counties.

Conditioning on county and year-fixed effects, we show that an additional hacking incident in a local organization leads to 1.044 more cases of identity theft per 10,000 population in the county in the following year. On the other hand, conditioning on CBSA and year fixed effects, an additional hacking incident leads to 0.792 more cases of identity theft per 10,000 population in the following year. The increases in identity theft are both statistically and economically significant. The increases represent a 42% (for CBSA) to 77% (for counties) hike from the average prevalence of identity theft.

The higher marginal impact of a hacking attack on identity theft in a county than in a CBSA may seem to suggest that for identity theft the distance effect dominates the urbanization effect. We should however remark that CBSAs have larger populations than counties; the mean county and CBSA populations for our samples are 96,166 and 277,837, respectively. Therefore, the marginal impact of an additional hack is to increase identity theft by 10 in a county and 22 in a CBSA. We confirm that urbanization play a significant role in determining the marginal effect

of a hacking incident on identity theft, by dividing the county sample into urban and rural counties and rerunning our analysis.

A large economics literature has studied the determinants and consequences of crime, both theoretically and empirically. This paper belongs to the growing, but much smaller, subset of this literature that focuses on cybercrime and cybersecurity. The theoretical contributions in this literature have focused on the network nature of cyberspace (e.g., see, Acemoglu et al., 2016). In particular, Roberds and Schreft (2009) analyze an environment where agents join payment networks to facilitate trade; because of the presence of positive network externalities in data management and the risk of breaches, there is socially excessive collection of personal information and too little security spending.

The empirical contributions to the economics literature on cybercrime have been mostly qualitative and descriptive (e.g., see Schreft, 2007; Moore et al., 2009; Kshetri, 2010). This is in part because empirically linking data breaches and financial damages is challenging, due to the technical nature of cyberattacks and the difficulties of attributing personal losses to specific breaches. For example, many victims of identity theft do not know how their information was stolen. Nevertheless, given the amount and detail of the personal information stored in modern computer systems and the scope of the digital economy, it is reasonable to suspect that data breaches contribute to fraud such as identity theft (Anderson et al., 2008). This paper contributes to the literature by empirically establishing that there is indeed a causal relationship between data breaches of a local organization and the subsequent increase of identity theft in the local area, and by providing a lower bound on how extensively individuals are victimized.

This paper is also related to the emerging literature on the economics of privacy (e.g., see Acquisti et al., 2013, for a review). An empirical challenge in this literature has been the difficulty to measure the trade off between the benefits of the modern digital economy and the costs associated with the loss of privacy (Acquisti et al., 2013; Athey et al., 2017). There is also a large criminology, sociology, and psychology literature revolving around the demographic and social predictors of victims/offenders (Gordon, Rebovich and Choo, 2007; Harrell, 2021; Anderson, 2006; Harrell, 2019), the resulting aftermath of victimisation (Reynolds, 2021; Golladay and Holtfreter, 2017), and the emotional consequences of identity theft (Golladay and Holtfreter, 2017).

The rest of the paper is structured as follows. Section 2 outlines the data. Section 3 describes the empirical strategy and specifications. In Section 4, we report and discuss the main findings. Section 5 provides some extensions and robustness checks. Section 6 concludes and suggests possible future extensions of our research.

2 Data and Descriptive Evidence

2.1 Hacking

We obtained hacking data from the Privacy Rights Clearinghouse (PRC) from 2015 to 2018. PRC has compiled a chronological database of publicly announced data breaches of private and public sector organizations in the USA. Their data is mainly sourced from the attorney general, media reports, US Department of Health and Human Services, and IT websites (e.g., databreaches.net, Krebs on security).[3] The database also provides a plethora of information linked to eight types of data security incidents across eight different sectors. Out of those eight categories of breaches, we confine our analysis to hacking-related data.

PRC defines a hacking incident to be an instance where the organization's information system is hacked by an outside party or infected with malware (Privacy Rights Clearinghouse, 2022). Hence, all breaches categorized as a "HACK" occur due to cybersecurity attack methods such as phishing, spoofing, and ransomware. We pin down the location of the hack to the locality of the breached organization. Next, we match the relevant county FIPS and the CBSA code to the locality where the breach occurred. Finally, we collapse the data at the county×year, or CBSA×year, level to obtain the total hacks and its first lag - the main variables of interest.

We note two limitations of our data. First, it doesn't allow us to identify the exact geographical spread of a hacking incident. To minimize the impact of this, we exclude a subset of observations (around 40%) that we consider to be "global hacks". These are the hacks of large private and state organizations that have headquarters in one county but carry out business operations beyond the boundaries of a county or CBSA. For instance, cyber attacks on social media such as Facebook, Twitter, chain stores, online retail stores that deliver across a number of states, and

[3] Please refer to Table A.1 of the Appendix for a list of sources related to our sample years.

international hotel chains are some of those global hacks that we exclude in the final hacking sample. We do so because, although we could pin down these hacks to their headquarters, we would not know the specific geographic locations of the potential victims of the breach. Thus, including these hacks would introduce substantial measurement error. Second, the PRC is one of the few databases that provide a large collection of data breaches covering the whole USA. It provides reliable, timely, and historical information, but fails to provide a complete picture of the data breaches in the USA. This is due to the discrepancies in data breach notification laws among states and in the different practices of the states' attorney generals. State law mandates the minimum number of victims for which organizations must report the data breach to the attorney general (Coie, 2021; Romanosky et al., 2011), but the majority of the attorney generals and state governments do not publish the breach notifications, making it difficult to build a comprehensive picture (Clearinghouse, 2018).[4] Because of this limitation, we are likely to be underestimating the impact of a local hacking incident on identity theft.

2.2 Identity Theft

We use identity theft data from the National Incident-Based Reporting System (NIBRS) of the FBI Uniform Crime Reporting (UCR) Program from 2016-2018. Relative to the traditional Summary Reporting System (SRS), NIBRS provides a more comprehensive picture of the crime scene in the USA. Also, it covers a wide spectrum of information related to offense categories, arrestees, and demographics of victims as well as offenders. One of the main advantages of this database is that it is a compilation of the official crime statistics of the USA reported by the law enforcement agencies (LEA) spread across the nation. They categorise all LEAs under several layers of jurisdictions (city, county, state police, tribal, educational institutions, or other). Hence this database allows greater flexibility to aggregate crime data starting from the smallest geographical unit (Strom and Smith, 2017). As of 2018, 7283 agencies submitted data via NIBRS, that approximately covers 36% of

[4] The minimum required by the law of each state is as follows: 50 (District of Colombia), 250 (North Dakota, Oregon, South Dakota, Texas), 500 (California, Colorado, Delaware, Florida, Illinois, Iowa, Rhode Island, Washington), 1000 (Alabama, Alaska, Arizona, Arkansas, Hawaii, Kansas, Kentucky, Maryland, Missouri, New Mexico, South Carolina, Virginia) and 10,000 (Georgia) victims or more. The remaining states have no minimum specified.

the US population. According to Strom and Smith (2017), a greater fraction of data represents the medium and smaller agencies associated with suburban and rural jurisdictions. While it is difficult to derive national-level estimates, we can utilize the local-level data to identify within-jurisdiction variation in the crime levels.

For our analysis, we construct two samples to cover two geographical layers, one at the county level and another at the CBSA level. Using the World Cities Database of Simple Maps (`https://simplemaps.com/data/world-cities`), we match the county or CBSA name/code to each agency included in the identity theft database.[5] Similar to the hacking sample, we exclude the identity theft observations of agencies that cater beyond the local boundaries of a county or a CBSA. For example, data recorded by metro transit police, highway patrol, or federal agencies such as the United States Army are not included in the final identity theft sample. Next, we sum the data by their respective county or CBSA code in a given year to get the two identity theft samples. We restrict both samples to counties that reported identity theft data in all three years. Figure 1 displays the counties represented in the final sample.

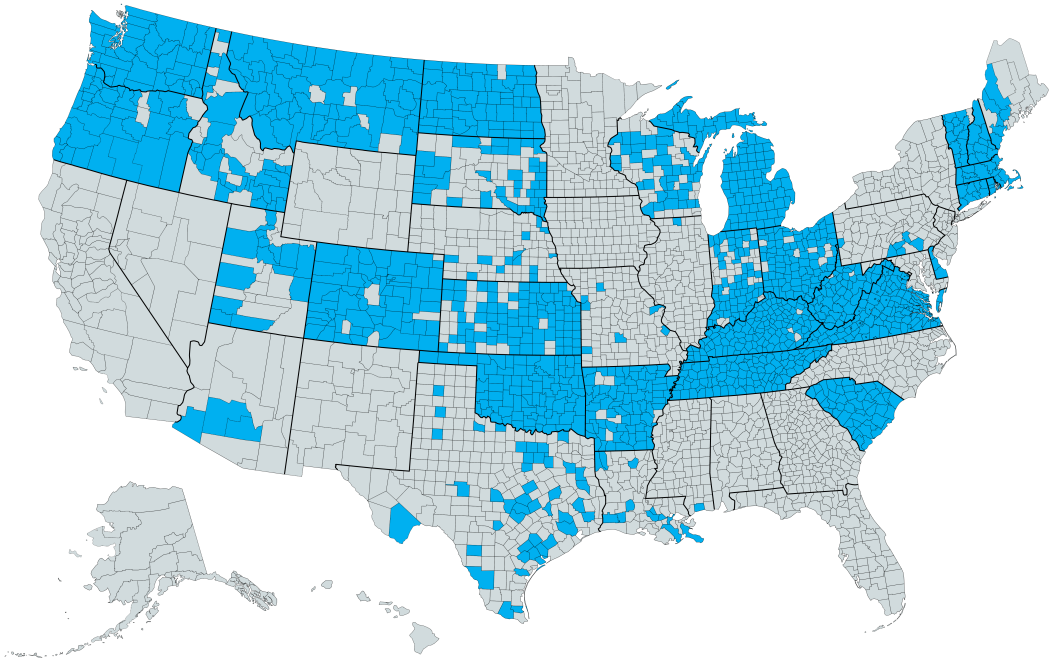
A limitation of our data is that the NIBRS captures only those identity theft incidents that are reported to the LEAs. Hence, it differs from survey data such as the National Crime Victimization Survey (NCVS), as it doesn't include unreported identity theft occurrences. People are more likely to report to the police if they know who their offender is, are communicating with a financial institution, have information related to the fraudulent activity that supports an investigation or it is a requirement by a creditor (Gerson, 2021; Reyns and Randa, 2017). Although including unreported identity theft would be valuable, the NCVS, providing a national representation of households, does not include comprehensive information about victimization in local geographical areas. Therefore, the NCVS survey data is not well suited to answer our research question.

2.3 Auxiliary Data

We obtained information related to unemployment from the U.S. Bureau of Labour Statistics (BLS). The Local Area Unemployment Statistics program of BLS provides

[5] We added county names/FIPS of certain townships that are not included in the world cities database before merging with the crime file.

Figure 1: The Counties Represented in the Final Sample



Note: The bold black lines indicates the state borders and the counties colored in blue represent the ones included in the final sample.

county-level estimates of the unemployment rate in the USA. These estimates are derived using "the handbook method": a method where the figures are calculated using a variety of sources including, but not limited to, current population survey and quarterly census of employment and wages (U.S. Bureau of Labor Statistics, 2018). Next, we collapse the total unemployment and labour force figures by their respective metropolitan or micropolitan statistical area to obtain the unemployment rate for the CBSA sample.

We also use the data breach notification laws (DBNL) compilation of Coie (2021) to capture the state level variation in the DBNL enactments or revisions. Accordingly, out of the 35 states represented in our final sample, South Dakota enacted the law in 2018, and 17 states made revisions to their existing laws during 2015 - 2018. Further, some changes made to the existing law include definition of personal information, timeline to notify the potential victims, and the content of the attorney general notification (Coie, 2022). We create a binary contemporary and lagged variable taking a value of 1 if a state has DBNL enactment or revision.

2.4 Main Samples

To obtain the final samples necessary for our analysis, we merge the hacking data with the identity theft data. The main sample consists of 1390 counties/county equivalents.[6] The second sample consist of 846 of the counties, aggregated into 452 CBSAs; thus, 544 counties don't belong to the CBSA sample. The United States Office of Management and Budget (OMB) defines a CBSA as a geographical entity with at least one core area consisting of 10,000 or more population and adjacent territories that have a high level of economic and social integration with the core in terms of commuting ties.

The first reason to analyze both at the county and the CBSA level is that CBSAs are more urbanized than counties on average, and the degree of urbanisation is a significant predictor of crime in general. We thus expect that hackers would consider organizations in more densely populated areas to be better targets, allowing them to make more lucrative gains. Indeed, previous research has found that

[6] The USA Census Bureau considers all independent cities of Maryland, Missouri, Nevada, Virginia, and boroughs, municipalities, and census areas of Alaska to be equal to a county for statistical purposes. However, these county-equivalents are independent of a county and are part of the primary divisions of their state (United States Census Bureau, 2021).

identity theft is positively correlated with the fraction of the population in urban areas (Lane and Sui, 2010). Hence, it is interesting to see how a marginal hacking incident influences identity theft in a CBSA.

Second, the CBSA sample also allows us to capture any spillover effects of hacking on counties. Given that metropolitan/micropolitan areas consist of outlying counties that have strong commuting ties with the urban core, a hacking incident on a particular organization's database could include the personal information of customers/employees living in adjoining counties.

There are two main reasons to distinguish between CBSAs and counties. First, CBSAs are more urbanized than counties on average, and the degree of urbanization is a significant social predictor of crime in general. Because of this "urbanization effect" one might expect the impact of hacking on identity theft to be greater in CBSAs. Second, CBSAs are larger geographical areas, with greater average distance from the hacked organization. Since we focus on localized hacking attacks, because of this "distance effect" one might expect the impact of hacking on identity theft to be greater in counties.

Table 1 presents the summary statistics of the county and CBSA samples. As shown in Column (1), a county reports, on average, 1.359 identity theft incidents per 10,000 population in a year. A county, on average, experiences 0.039 hacking incidents per 10,000 population in a given year. Similarly, almost 3% of the counties record at least one hacking incidence, and 2.5% record only one hack in a particular year.

On the other hand, the statistics of Column (2) reflect higher mean values for both identity theft and hacking-related variables in CBSA. This is because collapsing data at the CBSA level removes 544 counties that lie outside the metro/micropolitan area.[7] Moreover, approximately 72% and 99% of those county-year observations have zero identity theft and hacking records, respectively. The comparison suggests that identity theft and hacking incidents are heavily concentrated within the CBSAs.

[7] The mean figures for the CBSA sample, including those counties, are marginally higher than the county sample.

Table 1: Descriptive Statistics for County and CBSA Samples

	County	CBSA
	(1)	(2)
Identity theft per 10,000 population	1.359	1.879
	(3.265)	(3.820)
Total hacking Incidents	.039	.117
	(.240)	(.482)
Total hacking incidents (t-1)	.025	.084
	(.179)	(.381)
Being Hacked	.031	.077
	(.174)	(.267)
Being Hacked (t-1)	.022	.061
	(.147)	(.240)
One hacking Incident	.025	.053
	(.157)	(.224)
One hacking Incident (t-1)	.019	.046
	(.136)	(.209)
More than one hacking incident	.006	.024
	(.077)	(.154)
More than one hacking incident (t-1)	.003	.015
	(.056)	(.124)
Number of jurisdiction-year observations	4170	1356
Number of jurisdictions	1390	452

Note: This table reports the means and standard deviations (in parentheses) of variables for the county sample in Column (1) and the CBSA sample in Column (2). The outcome variable is identity theft per 10,000 population, based on data from the NIBRS of the UCR system. The hacking data are obtained from the Privacy Rights Clearinghouse, and population estimates for counties/metro and micropolitan statistical areas are obtained from the U.S. Census Bureau. All variables are aggregated at either the county or CBSA level. "Being hacked" is a binary indicator that takes a value of 1 if there is a hacking incident in county/CBSA *i* in year *t*. "Being hacked (t-1)" is a lag variable of "Being hacked", indicating hacking incidents in year *t-1*.

3 Empirical Approach and Specifications

We now turn to exploring the causal relationship between hacking of a local organization and identity theft at two different geographical layers, county and CBSA.

We begin by investigating the empirical link between hacking incidents and identity theft, controlling for various local economic factors. We first estimate the following equation:

$$\texttt{ID}_{i,t} = \alpha + \beta \cdot \texttt{HACK}_{i,\tau} + \gamma \cdot \texttt{X}_{i,t} + \delta_t + \mu_i + \varepsilon_{i,t} \tag{1}$$

where $\texttt{ID}_{i,t}$ is the number of identity theft incidents per 10,000 population in i at year t, where i represent a county or CBSA; $\texttt{HACK}_{i,\tau}$ is a vector of the hacking measures of county/CBSA i at year τ, where τ is equal to either t or $t-1$. The two measures of hacking incidents are the total hacking incidents and a binary indicator equal to one if there is a hacking incident in i at year τ. $\texttt{X}_{i,t}$ is a vector of county and state level controls: Unemployment rate, data breach notification law at year τ, population growth, property crimes and total police officers per 10,000 population. Finally, δ_t are year fixed effects, μ_i are county or CBSA fixed effects, and $\varepsilon_{i,t}$ is the error term.

Next, we explore the incremental effects of "n" hacking incidents on identity theft. Multiple hacking attempts aimed at different local firms across a given county allow the hackers to obtain the personal information of a greater number of individuals. Furthermore, in some cases, these hackers merge incomplete information received from different sources to form a complete set of personal information for each individual (Schreft, 2007; Ablon, 2018). For example, if individual A's health records in the local hospital include name, date of birth, and social security number and her online account at a local supermarket contains bank account details, merging the two gives a complete array of personal information.

Hence, we generate three binary variables, that is, "zero hacks," "one hack," and "more than one hack." The specification, in this case, is the following:

$$\texttt{ID}_{i,t} = \alpha + \beta_1 \cdot \texttt{ONE HACK}_{i,\tau} + \beta_2 \cdot \texttt{MORE THAN ONE HACK}_{i,\tau} + \gamma \cdot \texttt{X}_{i,t} + \delta_t + \mu_i + \epsilon_{i,t} \tag{2}$$

where $\texttt{ONE HACK}_{i,\tau}$ is a binary variable that indicates whether a county/CBSA had only one hacking incident at year τ; $\texttt{MORE THAN ONE HACK}_{i,\tau}$ is a binary indicator that takes a value of 1 if a county/CBSA had more than one hacking incident at year τ,

where τ is equal to either t or $t-1$. The other variables are defined as before in Equation (1).

We include the first lag of the hacking variables in our specifications for two reasons. First, data breaches can go unnoticed by firms and victims from a couple of days to several years. It takes time for a firm to realise that there is a security breach to their data servers and further investigation is often needed before the affected individuals can be notified. In most instances, people realise that they are a victim of identity theft after being contacted by a financial institution such as a bank or a credit card company (Harrell, 2021, p.7). According to the identity theft survey of US adults conducted by Synovate (2007), victims of new accounts and other frauds are more likely to identify the misuse of their personal information at least six months after the initial occurrence of the misuse than existing credit card and other existing accounts.

Moreover, the survey results of Benner, Givens and Mierzwinski (2000) indicate that the average duration of misuse is 14 months. In addition, according to the recent survey conducted by Identity Theft Resource Center (2021), approximately 36 percent of respondents took six months or more to discover that they were victims. Second, identity theft due to hacking incidents occurring in the later part of a given year is likely to be reported in the following year.

We estimate the two-way fixed effect models specified in Equations (1) and (2) to implement a difference-in-differences strategy that exploits the variation in hacking incidents over time within an entity (either county or CBSA), controlling for unobserved time-invariant geographic characteristics and time fixed effects. We use clustered-robust standard errors, clustering by county or CBSA to allow autocorrelated errors within a geographic area.

Our identification rests on the assumption that hacked and unhacked areas share parallel trends in potential outcomes of identity theft. If the parallel trend assumption holds, our estimated βs have causal interpretation. We note that reverse causation is unlikely to bias our treatment estimates of the impact of hacking on identity theft. In particular, we do not expect identity theft at time t to affect the hacking incidents at $t-1$.

To test for parallel trends, we include binary variables to indicate the two periods leading to the hacking incidents (see., e.g, Hu et al., 2020; Mello, 2019).[8]

[8] The short nature of our panel limits our ability to test for common trends between the treatment

This method requires a staggered adoption for treatment, i.e., once the treatment is switched on, it remains active in all subsequent periods. Given that our treatment can switch on and off at different points in time, we can test for parallel trends in a subset of the data, comprising 88% - 95% of the original sample that follows a staggered adoption design. In other words, our sample consists of counties/CBSAs that had no hacking incident in all three years (control group) and counties/CBSAs that had a hacking incident in the final year (i.e., 2018) of the sample (treatment group).[9]

4 Main Results

Table 2 presents the results of estimating Equation (1) using the county sample. We provide two different sets of results for our independent variables. In Columns (1) to (3), the main independent variable is the total number of hacking incidents in year τ. In Columns (4) to (6), the main independent variable is a binary indicator that takes a value of 1 if there was any hacking incident in county i in year τ, where τ is equal to either t or $t-1$.

While the first set looks at the marginal impact of a hacking incident, the binary variable shows the expected difference in identity theft outcomes between counties with and without any hacking incident. In all instances, the coefficients of interest are positive, indicating a positive relationship between the hacking variables and

and the control groups over longer periods of time.

[9]Using the two pre-treatment periods (2016-2017) and one post-treatment period (2018), we estimate the following specification:

$$\texttt{ID}_{i,t} = \alpha + \delta_t + \beta_{-2} \cdot \texttt{D}_i \cdot \mathbb{1}(t = 2016) + \beta_{-1} \cdot \texttt{D}_i \cdot \mathbb{1}(t = 2017) + \beta_0 \cdot \texttt{D}_i \cdot \mathbb{1}(t = 2018) + + \gamma \cdot \texttt{X}_{i,t} + \delta_t + \mu_i + \epsilon_{i,t} \quad (3)$$

where $\texttt{D}_i$ is an indicator variable equal to zero if county/CBSA i had no hacking incidents between 2016 and 2018, and one if it had one or more hacking incidents in 2018; For $\tau = 2016, 2017, 2018$, $\mathbb{1}(t = \tau)$ is an indicator function that is equal to one if $t = \tau$, and zero otherwise. The remaining variables are the same as in Equation (1). Keeping β_{-2} as the reference group, β_{-1} captures any pre-treatment trend deviation between the treatment and control groups. Using this specification to check for parallel trends, the effect size and standard error of β_{-1} is 0.209 (-0.150) & 0.504 (0.522), respectively, for county(CBSA) analysis. We do not find statistically significant trend deviations between the treatment and control groups. We cannot reject the null hypothesis that $\beta_{-1} = 0$ at the 10% level of significance. However, the contemporary treatment coefficient β_0 cannot be precisely estimated. The estimate of β_0 is -0.215 (-0.648) with a standard error of 0.548 (0.705) in the county (CBSA) sample. This is likely due to the much smaller sample of hacking incidents, and as shown in Section 4, the delayed impacts of hacking incidents on identity theft.

Table 2: Effect of Hacking on Identity Theft - Baseline Estimates (County Level)

	(1)	(2)	(3)	(4)	(5)	(6)
	Identity theft per 10,000 population					
Total Hacking Incidents	0.934**	0.370	0.415			
	(0.400)	(0.371)	(0.369)			
Total Hacking Incidents $(t-1)$	1.289**	1.060**	1.044**			
	(0.535)	(0.417)	(0.412)			
Being Hacked				1.185**	0.693	0.765
				(0.565)	(0.513)	(0.510)
Being Hacked $(t-1)$				1.479***	1.355***	1.320***
				(0.562)	(0.465)	(0.462)
Unemployment Rate			0.341***			0.342***
			(0.071)			(0.071)
Data Breach Notification Law			0.438***			0.438***
			(0.131)			(0.131)
Data Breach Notification Law $(t-1)$			-0.400***			-0.402***
			(0.102)			(0.102)
Property Crimes per 10,000 Population			0.005***			0.005***
			(0.002)			(0.002)
Total Police Officers per 10,000 Population			0.010			0.010
			(0.019)			(0.019)
Population Growth Rate			-0.042			-0.042
			(0.042)			(0.042)
R-squared	0.039	0.109	0.145	0.037	0.110	0.146
Number of Observations	4170	4170	4170	4170	4170	4170
Number of Counties	1390	1390	1390	1390	1390	1390
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
County Fixed Effects	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Robust standard errors clustered by county are in parentheses. The outcome variable is the number identity theft per 10,000 population, as the column headers describe. The independent variables are total hacking incidents, binary indicators taking a value 1 if any hacking incident occurred (Being Hacked) and their respective first lags. All regression models exclude hacks that are not at the local level.

identity theft per 10,000 population. However, given the time lags for firms and individuals to identify the potential breach, we expect a substantial proportion of breaches - particularly the hacking incidents of the second half of the year - to be reported in the year following the occurrence. Indeed, the regression outcomes indicate a higher value and greater statistical significance (at 1% - 5%) for the one-year lagged variable relative to contemporary variables.

In each set of predictor variables mentioned above, the first column provides the simple correlation between hacking measures and identity theft per 10,000 population after controlling only for year fixed effects. As we add the county fixed effects, as shown in Columns (2) & (5), the contemporary variables become statistically insignificant, yet the coefficient remains positive. On the other hand, the statistical significance of the lagged variables remains intact at 1 or 5 percent. The point estimates become more precise and have smaller standard errors.

Our results remain consistent even after controlling for local economic conditions, as well as other local and state-level control variables. In the full model, one hacking incident in a county in a given year leads to an average of 1.044 identity theft incidents per 10,000 population in the following year. Considering the sample mean of 1.359, such an increase translates to approximately a 77% increase in the rate of identity theft. Similarly, compared to counties with no hacking incidents, counties that experience hacking incidents, on average, see an additional 1.320 identity thefts per 10,000 population, which is a 97% increase from the mean rate of identity theft.

For comparison, we also report in Table A2 estimates from (i) the pooled ordinary least squares in Column (1), (ii) the random effects model in Column (2), and (iii) the fixed effect model in Column (3). First, the Breusch-Pagan Lagrange multiplier (LM) test rejects pooled OLS in favor of a random effect specification at a 1% significance level. Second, given that the standard Hausman test is invalid with the clustered-robust standard errors (Cameron and Trivedi, 2005), we perform a Sargan-Hansen test to compare between RE and FE models. As reported in the last Column of Table A2, we strongly reject the null hypothesis, indicating that time-invariant county-specific fixed effects correlate with the independent variables.

Table A2 uses the county sample. Similarly, we present the results for the CBSA sample in Table A3. Under both tests, we reject the null hypotheses that favor pooled OLD or RE models. Together, these results demonstrate that the fixed effect

model is the most appropriate for our county and CBSA analysis. Therefore, we focus on the fixed effect specification throughout this paper.

Next, we estimate Equation (1) using the CBSA sample and present the results in Table 3. Similar to the county-level analysis, we include the same set of independent variables, year fixed effects, CBSA fixed effects, and other control variables. The main results are similar to those reported in the county-level analysis in Table 2. As shown in Column (3), the marginal effect of a lagged hacking incident on identity theft per 10,000 population is approximately 0.792 for a CBSA. As illustrated in Column (6), a CBSA that experienced a hacking incident, in comparison to one with no such event this year, on average, sees an increase of 1.517 in identity theft incidents per 10,000 population in the following year. Given the sample mean of 1.879, these marginal effects represent, respectively, an approximately 42% and 81% increase from the average identity theft rate for CBSA.

Two main factors likely contribute to the difference in the impact of a hacking incident at the county and CBSA levels: urbanisation and distance from the hacked entity. Since the CBSA sample includes only counties that are part of a metro or micropolitan statistical area, we expect greater urbanisation to lead to a larger effect of a hacking incident at the CBSA level. Conversely, working in the opposite direction and potentially lowering the treatment effect at the CBSA level, is the distance effect. There is spatial decay in the impact of a hack at the CBSA level, due to the greater average distance between a hacked entity and potential identity theft victims. Comparing Table 3 with Table 2, we observe that the marginal effect of a hacking incident is lower in a CBSA. This seems to suggest that the spatial decay effect outweighs the urbanisation effect at the CBSA level.

To check for an urbanisation effect, we run the county analysis by dividing the sample into two categories: urban and rural counties. The urban counties are the same metro/micropolitan counties included in the CBSA sample. We present the outcomes in Table 4. Columns (1) and (3) show the estimates for the urban counties, while Columns (2) and (4) present the estimates for the rural counties.

Comparing Columns (1) and (3) with Columns (2) and (4), we observe that urbanisation plays a significant role in determining the marginal effect of a hacking incident on identity theft. The effect size for the rural counties is positive but smaller than that for urban counties. However, the estimates from the rural counties are not precise. None of the key hacking variables has a statistically significant

Table 3: Effect of Hacking on Identity Theft - CBSA Level

	(1)	(2)	(3)	(4)	(5)	(6)
	Identity theft per 10,000 population					
Total Hacking Incidents	0.228	0.228	0.210			
	(0.240)	(0.308)	(0.304)			
Total Hacking Incidents $(t-1)$	0.519	0.827**	0.792**			
	(0.345)	(0.340)	(0.329)			
Being Hacked				0.693	0.792	0.773
				(0.478)	(0.576)	(0.573)
Being Hacked $(t-1)$				1.013**	1.584***	1.517***
				(0.471)	(0.517)	(0.505)
Unemployment Rate			1.045***			1.037***
			(0.198)			(0.197)
Property Crimes per 10,000			0.007**			0.007**
Population			(0.003)			(0.003)
Total Police Officers per			-0.025			-0.024
10,000 Population			(0.057)			(0.057)
Population Growth			-0.051			-0.057
			(0.127)			(0.127)
R-squared	0.047	0.176	0.217	0.051	0.182	0.223
Number of Observations	1356	1356	1356	1356	1356	1356
Number of CBSAs	452	452	452	452	452	452
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
CBSA Fixed Effects	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Robust standard errors in parentheses, clustered at CBSA level. The outcome variable is the identity theft per 10,000 population, as the column headers describe. The independent variables are total hacking incidents, binary indicators taking a value 1 if any hacking incident occurred and their respective first lags. All regression models exclude hacks that are not at the local CBSA level.

coefficient. There is low variation in hacking incidents at the rural level, since in our sample hacking predominantly occurred in urban areas (see Table 4 for more details). This makes it challenging to precisely determine the treatment effect in the rural analysis.

Among the controls included in our model equation, the unemployment rate is a statistically significant predictor of identity theft. The point estimates of Columns (3) and (6) of Table 2 indicate that a 1% increase in the unemployment rate is associated with around a 0.34 increase in the rate of identity theft; a 25% increase from the mean. This high predictive power may be driven by those urban areas with high rate of identity theft. Indeed, the results of Table 4 suggest that the relationship between unemployment and identity thieves is higher in urban, with both point estimates and level of statistical significance higher in the urban counties than in the rural counties.

Yet, unemployment seems to have a stronger effect on identity theft at the CBSA level than in urban counties. This may be because the labour market is not at the county level (Tolbert and Sizer, 1996); county lines are political and geographical delineations that may not adequately capture the local labour market conditions. Since some people may commute daily across county boundaries, commuting zones (CZ) are a more meaningful representation of labour markets at the local level (Autor, Dorn and Hanson, 2013). In our analysis, unemployment at the CBSA level is likely a better representation of the local labour market conditions since each CBSA consists of an urban central core and neighbouring counties with which it has strong commuting ties. Hence, the county unemployment rate is a noisy measure that may bias the effect size towards zero.

Table 4: Hacking & Identity Theft - Urban vs. Rural Counties

	(1) Urban	(2) Rural	(3) Urban	(4) Rural
	Identity theft per 10,000 population			
Total Hacking Incidents	0.406	-0.048		
	(0.382)	(0.184)		
Total Hacking Incidents $(t-1)$	0.989**	0.576		
	(0.414)	(0.531)		
Being Hacked			0.790	-0.048
			(0.538)	(0.184)
Being Hacked $(t-1)$			1.261***	0.576
			(0.468)	(0.531)
Unemployment Rate	0.658***	0.095	0.660***	0.095
	(0.143)	(0.066)	(0.144)	(0.066)
Data Breach	0.452**	0.250*	0.454**	0.250*
Notification Law	(0.192)	(0.146)	(0.192)	(0.146)
Data Breach Notification	-0.560***	-0.291**	-0.563***	-0.291**
Law (t-1)	(0.162)	(0.126)	(0.162)	(0.126)
Property Crimes	0.006**	0.005*	0.006**	0.005*
per 10,000 Population	(0.003)	(0.003)	(0.003)	(0.003)
Total Police Officers	-0.021	0.032**	-0.021	0.032**
per 10,000 Population	(0.036)	(0.016)	(0.036)	(0.016)
Population Growth Rate	-0.084	0.019	-0.085	0.019
	(0.064)	(0.055)	(0.064)	(0.055)
R-squared	0.173	0.118	0.174	0.118
Number of Observations	2538	1632	2538	1632
Number of Counties	846	544	846	544
Year Fixed Effects	Yes	Yes	Yes	Yes
County Fixed Effects	Yes	Yes	Yes	Yes

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Robust standard errors clustered by county are in parentheses. The outcome variable is the number identity theft per 10,000 population. The key independent variables are total hacking incidents, binary indicators taking a value 1 if any hacking incident occurred (Being Hacked), and their respective first lags. The odd and even columns show the urban and rural county analysis, respectively.

Finally, to assess the complementarity in personal information, we present the results of estimating Equation (2) in Table 5. The first and second columns show the regression outcome based on the county and CBSA samples, respectively. As before, "One Hacking Incident" is a binary variable taking value 1 if the county/CBSA has only one hacking incident in year t. We also include the lag value of "One Hacking Incident" . There are few observations with more than two hacking incidents. Thus, we add a binary variable that takes a value of 1 if more than one hacking incident took place in the county/CBSA. Similarly, we include the lag of this binary indicator.

Having only one hacking incident in year $t-1$ increased the rate of identity theft in year t by 1.116 in a county and 1.539 in a CBSA. Having more than one hacking incident in a year $t-1$ increases the identity theft rate by 2.447. However, the additional impact is not precisely estimated and is not statistically significant at the 10% level.

On the other hand, for a CBSA, having more than one hacking incident in $t-1$ increases the identity theft rate by 1.439, which is significant at the 10% level. The point estimate is slightly smaller than the impact of the first hacking incident. However, we note that CBSA tends to be larger than a county, which means that two hacking incidents may be less likely to lead to leaks of overlapped personal information.

Therefore, there is some suggestive evidence of the incremental effects of multiple hacking incidents—accessing more than one database may allow the hackers to obtain more detailed personal information on the potential victim and make identity theft more valuable and, hence, more likely.

Table 5: Effect of the Number of Hacking Incidents on Identity Theft

	(1) County Sample	(2) CBSA Sample
	Identity theft per 10,000 population	
One Hacking Incident	0.740*	0.846
	(0.448)	(0.542)
One Hacking Incident $(t-1)$	1.116***	1.539***
	(0.388)	(0.500)
More than One Hacking Incident	0.780	0.558
	(1.048)	(0.810)
More than One Hacking Incident $(t-1)$	2.447	1.439*
	(1.581)	(0.854)
R-squared	0.147	0.223
Number of Observations	4170	1356
Number of Jurisdictions	1390	452
Jurisdiction Fixed Effects	County	CBSA
Year Fixed Effects	Yes	Yes
Controls	Yes	Yes

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Robust standard errors in parentheses are clustered at the county level in Column (1) and at the CBSA level in Column (2). The outcome variable is the identity theft rate per 10,000 population, as the column headers describe. "One Hacking Incident" is a binary indicator that takes a value of 1 if there is only one hacking incident in county/CBSA *i* in year *t*. "One Hacking Incident (t-1)" is a lag variable of "One Hacking Incident," indicating hacking incidents in year *t-1*. "More than One Hacking Incident" is a binary indicator that takes a value of 1 if there is more than one hacking incident in county/CBSA *i* in year *t*. "More than One Hacking Incident (t-1)" is a lag variable of "More than One Hacking Incident," indicating multiple hacking incidents in year *t-1*.

5 Extensions and Robustness Checks

In this section we begin by considering the impact of hacks in the medical and finance sector on identity theft. We then perform some robustness checks of our analysis.

5.1 Sector Heterogeneity

The PRC database categorises each organization as one of seven types. These categories are finance and insurance services, retail businesses, other businesses, education institutions, government and military, medical (including medical insurance services), and non-governmental organizations. Thus, in principle, the data should allow us to identify how a marginal hacking incident in a particular industry affects the identity theft per 10,000 population. This is important, because the impact of a hacking incident on identity theft is likely to differ based on the type of organization.

Some industries are more susceptible to cyberattacks than others due to the depth of personal information included in their databases and the lags in upgrading the cybersecurity infrastructure (on outdated software and operating systems in the medical sector, see Kapoor and Nazareth, 2013). For instance, a plethora of personal health information (PHI) related to each individual is accessible at once from the electronic data archives of the healthcare industry. Furthermore, PHI contains demographic (name, date of birth, address), financial (account and credit card numbers and CCV), insurance details, and social security numbers that could lead to financial, medical, and other forms of identity theft (Johnson, 2009). Indeed, medical and finance organizations are the two most sought-after types of organizations by cyber criminals in the recent past (Western Governors University, 2021; Manship, 2022). According to Seh, Zarour, Alenezi, Sarkar, Agrawal, Kumar and Ahmad Khan (2020), the medical sector, followed by the financial sector, has experienced the largest spike in data breaches during the 2015-2019 period, with the highest proportion of those health security breaches being due to hacking incidents.

Consistent with the above observation, the majority of hacks in our sample occurred in the medical sector (37.85%), followed by hacks in the financial sector (

25.83%). In Table 6, we estimate 4 specifications where the total number of hacking incidents is replaced by the total number of hacking incidents in a specific sector.

As shown in Column (1) of Table 6, each additional hacking incident in the medical sector increases the rate of identity thieves per 10,000 population by 1.32. The estimate is statistically significant at the 10% level. However, the estimated effect of medical hacks is not substantially different from the non-medical hacks. This may be due to the greater heterogeneity of hacks in other sectors and, hence, noise in the data. Furthermore, the impact of a financial hack is large but not statistically significant. A relatively smaller proportion of hacks in the financial sector makes it difficult to estimate the coefficient with precision (high standard error). Similarly, Columns (3) and (4) show that hacks in educational institutions and businesses in year $t-1$ lead to increases in the rate of identity theft. The estimates for education and business hacks are smaller and statistically insignificant.

5.2 Breach Announcements

The results presented in Section 4 depend on assigning a hacking attack to the year the breach occurred. Since some breaches are not discovered the same year that they are announced, we consider here how the results would change if we assigned an attack to the year the breach is announced. We use the same county sample and collapse the data by the year the breach became public news. The results are presented in Table 6.

The point estimates of the lag independent variables are statistically significant at 1% after controlling for fixed effects and other factors. Thus, an extra hacking incident raises identity theft by 0.777 per 10,000 population: a 57% increase from the mean of the outcome variable. This effect size is slightly attenuated relative to the effect reported in Table 2. Public announcements are not the only way people realise that their personal information is being wrongfully utilised. In most instances, they learn that their identity information has been compromised when a credit card/ financial company contacts them. Hence, it is not surprising that the impact of the year of public announcement would have a lesser impact than when the hacking attack actually occurred.

We also check if our general results are robust to using an unbalanced panel. The results of the unbalanced panel presented in Table 8 show estimates that are

Table 6: Effect of Hacking Incidents on Identity Theft - By organization Type

	(1)	(2)	(3)	(4)
	Identity theft per 10,000 population			
Total Medical Hacking Incidents $(t-1)$	1.322*			
	(0.678)			
Total Financial Hacking Incidents $(t-1)$		1.620		
		(1.187)		
Total Education Hacking Incidents $(t-1)$			0.095	
			(0.765)	
Total Business Hacking Incidents $(t-1)$				0.636
				(0.444)
Number of Observations	4170	4170	4170	4170
Number of Counties	1390	1390	1390	1390
Year Fixed effects	Yes	Yes	Yes	Yes
County Fixed Effects	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Robust standard errors in parentheses, clustered at county level. The outcome variable is the identity theft per 10,000 population, as the column headers describe. The independent variables are total hacking incidents of medical, financial, education and business sectors respectively at time $t-1$.All regression models exclude hacks that are not at the local level. Crime data comes from the National Incident-Based Reporting system and hacking data from the Privacy Rights Clearinghouse. The population estimates for counties are taken from the U.S. Census Bureau.

Table 7: Effect of Hacking Incidents on Identity Theft - Based on the Year the Breach was Made Public

	(1)	(2)	(3)	(4)	(5)	(6)
	Identity theft per 10,000 population					
Total Hacking Incidents	1.083**	0.424	0.459			
	(0.435)	(0.312)	(0.308)			
Total Hacking Incidents $(t-1)$	0.915	0.760***	0.777***			
	(0.576)	(0.283)	(0.280)			
Being Hacked				1.477***	0.679*	0.753*
				(0.554)	(0.395)	(0.390)
Being Hacked $(t-1)$				1.295*	0.974***	0.989***
				(0.678)	(0.359)	(0.360)
Unemployment Rate			0.345***			0.348***
			(0.072)			(0.072)
Data Breach Notification Law			0.441***			0.440***
			(0.131)			(0.131)
Data Breach Notification Law $(t-1)$			-0.396***			-0.397***
			(0.102)			(0.102)
Property Crimes per 10,000 Population			0.005***			0.005***
			(0.002)			(0.002)
Total Police Officers per 10,000 Population			0.010			0.010
			(0.019)			(0.019)
Population Growth Rate			-0.042			-0.043
			(0.042)			(0.042)
R-squared	0.037	0.106	0.142	0.036	0.106	0.143
Number of Observations	4170	4170	4170	4170	4170	4170
Number of Counties	1390	1390	1390	1390	1390	1390
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
County Fixed Effects	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Robust standard errors in parentheses, clustered at county level. The outcome variable is the identity theft per 10,000 population. The independent variables are total hacking incidents, binary indicators taking a value 1 if any hacking incident occurred and their respective first lags.

Table 8: Effect of Hacking on Identity Theft - Unbalanced Panel of County Sample

	(1)	(2)	(3)	(4)	(5)	(6)
	Identity theft per 10,000 population					
Total Hacking Incidents	1.031***	0.426	0.422			
	(0.345)	(0.347)	(0.341)			
Total Hacking Incidents $(t-1)$	1.235***	1.046***	1.033***			
	(0.492)	(0.380)	(0.372)			
Being Hacked				1.314**	0.727	0.721
				(0.546)	(0.509)	(0.507)
Being Hacked $(t-1)$				1.528***	1.377***	1.343***
				(0.538)	(0.453)	(0.449)
Unemployment Rate			0.376***			0.377***
			(0.064)			(0.065)
Data Breach Notification Law			0.552***			0.551***
			(0.124)			(0.124)
Data Breach Notification Law $(t-1)$			0.507***			0.502***
			(0.128)			(0.127)
Property Crimes per 10,000 Population			0.005***			0.005***
			(0.002)			(0.002)
Total Police Officers per 10,000 Population			0.011			0.011
			(0.014)			(0.014)
Population Growth Rate			-0.033			-0.033
			(0.040)			(0.040)
R-squared	0.040	0.103	0.132	0.037	0.103	0.132
Number of Observations	4790	4790	4790	4790	4790	4790
Number of Counties	1760	1760	1760	1760	1760	1760
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
County Fixed Effects	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Robust standard errors clustered by county are in parentheses. The outcome variable is the number identity theft per 10,000 population. The independent variables are total hacking incidents, binary indicators taking a value 1 if any hacking incident occurred (Being Hacked) and their respective first lags.

similar to our balanced county analysis in Table 2.

5.3 Heterogeneity of Treatment Effects

We obtained our causal estimates under two assumptions: parallel trends and constant treatment effects over time and across different U.S counties. However, it is possible that the impact of a hacking incident on identity theft does not remain constant across counties and changes over time. Hence, it is crucial to see if our results are robust to heterogeneous treatment effects.

In the case of two-way fixed effect regressions with two treatment variables, each treatment coefficient comprises the sum of two components. The first component captures the weighted sum of the treatment effect in each county i and time t. Since some of the weights attached to treated cells can be negative, having more and larger negative values could lower the robustness of the estimates to heterogeneous effects.

The second component is the weighted sum of the effect of other treatment variables on the first treatment. Thus, if the effect of these other treatments is non-homogeneous, it will contaminate the coefficient of the first treatment variable (de Chaisemartin and D'Haultfoeuille, 2022).

When considering total hacking incidents per 10,000 population (THI) as the treatment variable, the first component of the THI coefficient is the weighted sum of the effects of increasing the total hacking incidents by one per 10,000 population in 130 county × year cells. Of these cells, only l6 cells received a negative weight. Furthermore, the respective positive and negative weights sum to 1.067 and -0.067. Table 9 shows the comparable results when the treatment variable is lagged THI, or one of the other set of treatment variables (Being Hacked, Being Hacked $(t-1)$). The results of all four treatment estimates are similar in terms of having few and smaller negative weights. Hence, it is likely that our estimates are robust to heterogeneity effects.

In Table 10, we present the contamination weights of the second component of each treatment coefficient. For example, the second component of the THI coefficient reflects the sum of the effects of having a hacking incident in the previous year in 92 county × year observations. Note that 32% of the cells received positive weights, where those positive and negative weights sum to 0.198 and -0.198, respec-

Table 9: Analysis of the First Term of the Treatment Coefficient

Treatment variable	No. of county × year cells receiving		Sum of weights	
	Positive Weights	Negative Weights	Positive	Negative
THI	114	16	1.067	-0.067
THI(t-1)	85	7	1.006	-0.006
Being Hacked	121	9	1.015	-0.015
Being Hacked (t-1)	87	5	1.001	-0.001

Table 10: Analysis of the Second Term of the Treatment Coefficient

Treatment variable	No. of county × year cells receiving		Sum of weights	
	Positive Weights	Negative Weights	Positive	Negative
THI	29	63	0.198	-0.198
THI(t-1)	55	75	0.442	-0.442
Being Hacked	32	60	0.193	-0.193
Being Hacked (t-1)	56	74	0.392	-0.392

tively. In each case, although more than half of the treated cells received negative contamination weights, the sum of those weights is very small. Hence, our treatment variables don't seem to be heavily contaminated by the other treatment.

6 Concluding Remarks

We present the first set of empirical estimates of the effect of hacking on identity crime. The fixed effect estimation results presented in the previous section indicate that one hacking incident at time $t-1$ in a county or CBSA will result in a 1.044 and 0.792 identity theft per 10,000 population at time t, respectively. As sufficient time is needed to identify the misuse of personal information, we observe statistically significant results only for the first period lag of the hacking variables. This effect size is large relative to the mean of identity theft (77% & 42%) but is not unreasonable. Due to the lack of comprehensive data on identity theft and hacking,

there is no directly comparable study. Yet, intuitively, such a strong causal effect is plausible, as many of the organizations in our sample are small to medium sized.

First, there is ample evidence that many small to medium size businesses (SMB) have become prey to cyberattacks. According to the latest research done by Identity Theft Resource Center (ITRC), more than half of the small firms in the USA have suffered from a data breach incident (Muncaster, 2021). Similarly, a survey conducted by the Australian Cyber Security Center (2020) suggests that 62% of the small and medium firm participants experienced a cyber security incident. The lack of IT resources, failure to identify their weakness in security practices, underestimation of the cyber risk of small organizations due to large media focus on massive data breaches are some of the factors that highlight the vulnerability of small organizations to data security attacks (Australian Cyber Security Center, 2020).

Second, the motives of the cyber hackers also play a role in justifying our results. According to Ablon (2018), these cyber actors can be categorised into four main types: Cyberterrorist, hacktivists, state-sponsored actors, and cybercriminals. The cyberterrorist and state-sponsored actors may work with a national interest in mind; hacktivists are typically motivated by a political, economic, and social cause such as embarrassing celebrities, pointing out the security issues of organizations that have little impact on identity theft. All three types are unlikely to target SMBs primarily. Out of the four types, cybercriminals are the ones motivated by the financial gain that may have the largest impact on local identity theft. Most literature points out that the highly developed black market for cybercrime and the lower barriers to entry motivates criminals to steal personal information for financial gains (Ablon, 2018; Schreft, 2007). Thus, it is plausible that SMBs are mostly targeted by criminals with an intention of extracting and using the individuals' personal information.

Out of the controls we have included in our regression analysis, unemployment seems to play a significant role in identity theft. The positive effect size could be due to two factors. First, there is still a sizable proportion of identity theft occurring through traditional methods. From the victim's perspective, survey studies have shown that the majority of people did not know how their identity had been stolen. But for the ones that did, a fraction was stolen through low-tech methods that require some local physical proximity between the victim and the offender (Harrell, 2021). Second, some unemployed people may seek employment in the

informal online crime market. There is substantial division of labour in this market, suggesting that specialised knowledge in hacking is no longer an important entry barrier (Goodchild, 2021).

Our results suggest that public subsidies or regulation may be needed to enhance the IT security systems of SMBs. Small organizations may be reluctant to spend enough on IT security due to the asymmetric information and externalities associated with the protection of consumer data. The information asymmetry arises when the consumers cannot identify which organizations are taking greater precautions to protect their personally identifiable information. Hence, their average willingness to pay a higher price is low, leading to too little security provided by the organizations. It is a grave problem for SMBs, as protecting internal systems against a malware attack is very costly.

On the other hand, network effects in the payment systems suggest that an organization is at risk even if they have an adequate security investment. Since many parties – banks, internet service providers, vendors – are involved in the whole payment process, a lack of protection by one party puts the entire system at risk. Thus, the presence of market failure in the data security market reinforces the need for government intervention at the local level (Schreft, 2007).

While our paper is one of the first to explore how intentional data breaches, such as hacking, lead to identity theft, there is room for more work in this area. Although our main source of identity theft – NIBRS data – provides a good picture of identity crime at the local level, its geographical coverage is incomplete. Since the FBI has taken steps that require all reporting agencies to provide crime statistics via NIBRS starting from 2021, it will become possible to obtain a more comprehensive picture of the reported crime statistics in the USA. Furthermore, given that identity theft generates a variety of identity-related frauds, such as financial, unemployment, and child identity fraud, it would be interesting to explore how data breaches impact these crimes.

References

Ablon, Lillian (2018) *Data thieves: The motivations of cyber threat actors and their use and monetization of stolen data*: RAND.

Acemoglu, Daron, Azarakhsh Malekian, and Asu Ozdaglar (2016) "Network security and contagion," *Journal of Economic Theory*, Vol. 166, pp. 536–585.

Acquisti, Alessandro, Leslie K John, and George Loewenstein (2013) "What is privacy worth?" *The Journal of Legal Studies*, Vol. 42, pp. 249–274.

Almond, Douglas, Xinming Du, and Alana Vogel (2020) "Russian Holidays Predict Troll Activity 2015-2017,"Technical report, National Bureau of Economic Research.

Amir, Eli, Shai Levi, and Tsafrir Livne (2018) "Do firms underreport information on cyber-attacks? Evidence from capital markets," *Review of Accounting Studies*, Vol. 23, pp. 1177–1206.

Anderson, Keith B (2006) "Who are the victims of identity theft? The effect of demographics," *Journal of Public Policy & Marketing*, Vol. 25, pp. 160–171.

Anderson, Keith B, Erik Durbin, and Michael A Salinger (2008) "Identity theft," *Journal of Economic Perspectives*, Vol. 22, pp. 171–192.

Anderson, Ross and Tyler Moore (2006) "The economics of information security," *science*, Vol. 314, pp. 610–613.

Athey, Susan, Christian Catalini, and Catherine Tucker (2017) "The digital privacy paradox: Small money, small costs, small talk,"Technical report, National Bureau of Economic Research.

Australian Cyber Security Center (2020) "Cyber Security and Australian Small Businesses,"Technical report.

Autor, David H, David Dorn, and Gordon H Hanson (2013) "The China syndrome: Local labor market effects of import competition in the United States," *American Economic Review*, Vol. 103, pp. 2121–2168.

Bana, Sarah, Erik Brynjolfsson, Wang Jin, Sebastian Steffen, and Xiupeng Wang (2022) "Human capital acquisition in response to data breaches," *Available at SSRN 3806060*.

Benner, Janine, Beth Givens, and Edmund Mierzwinski (2000) "Nowhere to turn: Victims speak out on identity theft," *CALPIRG/Privacy Rights Clearinghouse Report*.

Buzzard, John and Tracy Kitten (2021) "2021 Identity Fraud Study: Shifting Angles," URL: https://www.javelinstrategy.com/content/2021-identity-fraud-report-shifting-angles-identity-fraud, publisher: Javelin Strategy & Research.

Cameron, A Colin and Pravin K Trivedi (2005) *Microeconometrics: methods and applications*: Cambridge university press.

de Chaisemartin, Clément and Xavier D'Haultfoeuille (2022) "Two-way Fixed Effects and Differences-in-Differences Estimators with Several Treatments,"Technical report, National Bureau of Economic Research.

Clearinghouse, Privacy Rights (2018) "Data Breach Notification in the United States and Territories."

Coie, Perkins (2021) "Security Breach Notification Chart," September, URL: https://www.perkinscoie.com/en/news-insights/security-breach-notification-chart.html.

——— (2022) "2022 Breach Notification Law Update: State and Federal Requirements Continue To Evolve," URL: https://www.perkinscoie.com/en/news-insights/2022-breach-notification-law-update-state-and-federal-requirements-continue-to-evolve.html.

Federal Trade Commission (2021) "Consumer Sentinel Network: Data Book 2020,"Technical report.

Gerson, Emily Starbuck (2021) "Should You File a Police Report After Identity Theft?," April, URL: https://www.experian.com/blogs/ask-experian/should-you-file-a-police-report-after-identity-theft/.

Golladay, Katelyn and Kristy Holtfreter (2017) "The consequences of identity theft victimization: An examination of emotional and physical health outcomes," *Victims & Offenders*, Vol. 12, pp. 741–760.

Goodchild, Joan (2021) "Cybercrime 'Help Wanted': Job Hunting on the Dark Web," March, URL: https://www.darkreading.com/edge/cybercrime-help-wanted-job-hunting-on-the-dark-web.

Gordon, Gary R, D Donald J Rebovich, and Kyung-Seok Choo (2007) "Identity fraud trends and patterns," *Center for Identity Management and Information Protection, Utica College*.

Harrell, Erika (2019) "Victims of Identity Theft, 2016,"Technical report, U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

——— (2021) "Victims of Identity theft, 2018,"Technical report, U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

Hu, Jiafei, Rigissa Megalokonomou, and Haishan Yuan (2020) "How do parents respond to regulation of sugary drinks in child care? Evidence from California," *Journal of Economic Behavior & Organization*, Vol. 178, pp. 672–687.

Identity Theft Resource Center (2021) "2021 ITRC Consumer Aftermath Responses: Non-Pandemic Related,"Technical report.

Johnson, M Eric (2009) "Data hemorrhages in the health-care sector," in *International Conference on Financial Cryptography and Data Security*, pp. 71–89, Springer.

Kahn, Charles M and William Roberds (2008) "Credit and identity theft," *Journal of Monetary Economics*, Vol. 55, pp. 251–264.

Kapoor, Akshat and Derek L Nazareth (2013) "Medical data breaches: What the reported data illustrates, and implications for transitioning to electronic medical records," *Journal of Applied Security Research*, Vol. 8, pp. 61–79.

Kshetri, Nir (2010) *The global cybercrime industry: economic, institutional and strategic perspectives*: Springer Science & Business Media.

Lane, Gina W and Daniel Z Sui (2010) "Geographies of identity theft in the US: understanding spatial and demographic patterns, 2002–2006," *GeoJournal*, Vol. 75, pp. 43–55.

Manship, Ryan (2022) "The Top 6 Industries At Risk For Cyber Attacks - RedTeam Security," URL: `https://www.redteamsecure.com/blog/the-top-6-industries-at-risk-for-cyber-attacks`.

Mello, Steven (2019) "More COPS, less crime," *Journal of Public Economics*, Vol. 172, pp. 174–200.

Mikhed, Vyacheslav and Michael Vogan (2018) "How data breaches affect consumer credit," *Journal of Banking & Finance*, Vol. 88, pp. 192–207.

Moore, Tyler, Richard Clayton, and Ross Anderson (2009) "The economics of online crime," *Journal of Economic Perspectives*, Vol. 23, pp. 3–20.

Muncaster, Phil (2021) "Small Businesses Pay Up to $1M to Recover from Breaches," October, URL: `https://www.infosecurity-magazine.com/news/small-businesses-paying-1m-recover/`.

Prince, Jeffrey T and Scott Wallsten (2022) "How much is privacy worth around the world and across platforms?" *Journal of Economics & Management Strategy*, Vol. 31, pp. 841–861.

Privacy Rights Clearinghouse (2022) "Data Breach Chronology," URL: https://privacyrights.org/data-breaches.

Reynolds, Dylan (2021) "The differential effects of identity theft victimization: How demographics predict suffering out-of-pocket losses," *Security Journal*, Vol. 34, pp. 737–754.

Reyns, Bradford W and Ryan Randa (2017) "Victim reporting behaviors following identity theft victimization: Results from the national crime victimization survey," *Crime & Delinquency*, Vol. 63, pp. 814–838.

Roberds, William and Stacey L Schreft (2009) "Data breaches and identity theft," *Journal of Monetary Economics*, Vol. 56, pp. 918–929.

Romanosky, Sasha, Rahul Telang, and Alessandro Acquisti (2011) "Do data breach disclosure laws reduce identity theft?" *Journal of Policy Analysis and Management*, Vol. 30, pp. 256–286.

Schreft, Stacey L (2007) "Risks of identity theft: Can the market protect the payment system?" *Economic Review-Federal Reserve Bank of Kansas City*, Vol. 92, p. 5.

Seh, Adil Hussain, Mohammad Zarour, Mamdouh Alenezi, Amal Krishna Sarkar, Alka Agrawal, Rajeev Kumar, and Raees Ahmad Khan (2020) "Healthcare data breaches: insights and implications," in *Healthcare*, Vol. 8, p. 133, Multidisciplinary Digital Publishing Institute.

Strom, Kevin J and Erica L Smith (2017) "The future of crime data: The case for the National Incident-Based Reporting System (NIBRS) as a primary data source for policy evaluation and crime analysis," *Criminology & Public Policy*, Vol. 16, pp. 1027–1048.

Synovate (2007) "Federal Trade Commission- 2006 Identity Theft Survey Report,"Technical report.

Tolbert, Charles M and Molly Sizer (1996) "US commuting zones and labor market areas: A 1990 update."

TransUnion (2021) "COVID-19's current and future impact on household budgets, spending and debt," URL: https://content.transunion.com/v/consumer-pulse-us-q4-2021?_gl=1*13uyyk9*_ga*MTk0MDExMzUyNC4xNjQxMDk2NjY0*_ga_6D2F5M2DQK*MTY0MTA5NjgyMC4xLjAuMTY0MTA5NjgyMi4w.

United States Census Bureau (2021) "Terms and Definitions," October, URL: https://www.census.gov/programs-surveys/popest/guidance-geographies/terms-and-definitions.html.

U.S. Bureau of Labor Statistics (2018) "Local Area Unemployment Statistics," January, URL: https://www.bls.gov/opub/hom/lau/pdf/lau.pdf.

U.S Department of Justice (2018) "2019 National Incident-Based Reporting System User Manual," URL: https://www.njsp.org/ucr/pdf/forms/nibrs_user_manual_20180403.pdf.

Western Governors University (2021) "6 Industries Most Vulnerable to Cyber Attacks," URL: https://www.wgu.edu/blog/6-industries-most-vulnerable-cyber-attacks2108.html#close.

Appendix

Table A1: Sources of Hacking Data

Source	Year of Breach			
	2015	2016	2017	2018
California Attorney General	x	x	x	x
Databreaches.net*	x	x	-	-
Government Agency	x	x	x	x
Indiana Attorney General	x	x	x	x
Krebs on security*	x	x	x	-
Maryland Attorney General	x	x	x	x
Media	x	x	x	x
Security breach letters	x	x	x	x
US Department of Health and Human Services	x	x	x	x
Vermont Attorney General*	x	x	x	-

Notes: The sources denoted with an * have less than 10 observations that are included in the final samples. The hacking data comes from the Privacy Rights Clearinghouse.

Table A2: Effect of Hacking on Identity Theft at county level - OLS, RE and FE Results

	(1)	(2)	(3)
	Identity theft per 10,000 population		
Total Hacking Incidents	0.640*	0.522	0.415
	(0.380)	(0.330)	(0.369)
Total Hacking Incidents (t-1)	0.989**	1.104***	1.044**
	(0.503)	(0.422)	(0.412)
Number of Observations	4170	4170	4170
R-squared	0.118	-	0.145
Controls	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes
County Fixed Effects	No	No	Yes
Pooled OLS vs random effects			
Ho: Pooled OLS model			
p-value			0.000
Fixed vs random effects			
Ho: Random effects model			
Sargan-Hansen statistic			52.150
p-value			0.000

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Robust standard errors in parentheses, clustered at county level. The outcome variable is the identity theft per 10,000 population, as the column headers describe. The independent variables are total hacking incidents and its first lag. All regression models exclude hacks that are not at the local level. Crime data comes from the National Incident-Based Reporting system and hacking data from the Privacy Rights Clearinghouse. The population estimates for counties are taken from the U.S. Census Bureau.

Table A3: Effect of Hacking on Identity Theft at CBSA level - OLS, RE and FE Results

	(1)	(2)	(3)
	Identity theft per 10,000 population		
Total Hacking Incidents	0.158	0.184	0.210
	(0.219)	(0.235)	(0.304)
Total Hacking Incidents (t-1)	0.514	0.736**	0.792**
	(0.329)	(0.298)	(0.329)
Number of Observations	1356	1356	1356
R-squared	0.136	-	0.217
Controls	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes
CBSA Fixed Effects	No	No	Yes
Pooled OLS vs random effects			
Ho: Pooled OLS model			
p-value			0.000
Fixed vs random effects			
Ho: Random effects model			
Sargan-Hansen statistic			125.066
p-value			0.000

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Robust standard errors in parentheses, clustered at CBSA level. The outcome variable is the identity theft per 10,000 population, as the column headers describe. The independent variables are total hacking incidents, binary indicators taking a value 1 if any hacking incident occurred and their respective first lags. All regression models exclude hacks that are not at the local CBSA level. Crime data comes from the National Incident-Based Reporting system and hacking data from the Privacy Rights Clearinghouse. The population estimates for counties are taken from the U.S. Census Bureau and collapsed to obtain the population of a metropolitan/micropolitan area. For the CBSA analysis, the controls are calculated by summing the estimates of the counties included in the sample.

Table A4: Sample Representativeness of U.S Population

Panel A: NIBRS Crime Data Analysis (2016-2018)				
	Metropolitan	Micropolitan	Rural	Total
USA Population	838790933	81741714	54499435	975032082
Sample Population				
Crime-Balanced	339412778	37335314	24264164	401012256
Crime-Unbalanced	367820355	41383142	27624111	436827608
Share of US population living in	86.03%	8.38%	5.59%	
Share of Sample population living in				
Crime-Balanced	84.64%	9.31%	6.05%	
Crime-Unbalanced	84.20%	9.47%	6.33%	
Panel B: Hacking Data Analysis (2015-2018)				
	Metropolitan	Micropolitan	Rural	Total
USA Population	1114090164	108959308	72721604	1295771076
Hacking Population	334105245	1550918	234649	335890812
Share of US population living in	85.98%	8.41%	5.61%	
Share of Sample population living in	99.47%	0.46%	0.07%	

Notes: The population estimates for counties are taken from the U.S. Census Bureau and collapsed to obtain the population of a metropolitan/micropolitan and rural area. Panel A & B provide the sum of population estimates of the counties that reported identity crime and hacking, respectively.