JOURNAL OF
EMERGING INVESTIGATORS

Article

Understanding the battleground of identity fraud

Sidhartha Basu[1], D'Arcy P. Mays[2]

[1] Deep Run High School, Glen Allen, Virginia

[2] Statistical Sciences and Operations Research, Virginia Commonwealth University, Richmond, Virginia

SUMMARY

Identity fraud has rapidly expanded into one of the fastest-growing white-collar crimes in the US. Beyond its substantial economic toll, resulting in billions of dollars in losses to the economy, identity fraud inflicts significant financial losses and mental distress upon its victims. With increasing online activity and the growing frequency of significant data breaches, the complexity and scale of identity fraud continue to grow. However, most academic research in this space has been focused on identifying cognitive behaviors and interventions at the individual level. Addressing a complex, multifaceted social issue like identity fraud necessitates a more comprehensive understanding of its underlying drivers at a broader macro level. We employed statistical methodologies to examine and analyze the factors influencing identity fraud in the US across a wide spectrum of variables using data from 2005 to 2021. A total of 12 explanatory variables, including macroeconomic indicators, sociodemographic factors, and criminal behavior, were analyzed. We identified the statistically significant variables associated with identity fraud through multiple linear regression, ANOVA, and multicollinearity analysis. Our analysis supported the hypothesis that the national unemployment rate, online banking usage, and incidence of fraud-related offenses were statistically significant variables in explaining identity fraud. Although not statistically significant, the increasing occurrence of data breaches and cyber-attacks and their implications for data security and privacy may warrant further attention. The overarching objective of this study was to establish a macro-level framework to understand identity fraud better, thereby fostering subsequent research and intervention efforts at both the individual and societal levels.

INTRODUCTION

In its broadest construct, financial fraud involves fraudsters using deceptive means to gain unauthorized access to an individual or an organization's monetary resources for their benefit. This can manifest in several ways, from imposter scams where fraudsters claim to be government representatives, friends, or distant family members in distress to scammers using stolen credit cards to make purchases on the victim's account. Identity fraud is another type of financial fraud where the identity of an existing person is used as a target or a principal tool without that person's consent for unlawful purposes (1). In addition to monetary losses that can be financially crippling, victims of identity fraud can face emotional and psychological impacts and feel complicit in the crime. According to data published by the Bureau of Justice Statistics in 2021, ten percent of identity theft victims reported being severely distressed because of the crime (2).

In the US, the Federal Trade Commission (FTC) has been collecting consumer-reported fraud data for several decades. Based on FTC Consumer Sentinel Network data, in 2021, there were 5.7 million reported fraud cases, resulting in 5.9 billion dollars in economic losses (3). Identity fraud made up 1.4 million of the total reported cases and is the fastest-growing fraud category, growing at an average rate of 46% year over year since 2017 (3). The large number of data hacks coupled with personal information being available in several places, such as phones, laptops, and public records, that fraudsters could steal from are making identity fraud increasingly sophisticated and hard to prevent (4). Most research has been done at the individual level and limited academic research has been done to understand the drivers of identity fraud at a macro level (5). Given the complex, multifaceted nature of identity fraud, it is worthwhile to further research at the macro level to understand the drivers of identity fraud better. Our study focuses on statistically understanding the drivers of identity fraud in the US using a broad canvas of macroeconomic, sociodemographic, and criminal behavior data from 2005 to 2021. This study aims to lead to a better understanding and identification of the drivers of identity fraud at a macro level that, in turn, will enable more precise and holistic interventions at the societal and individual levels to combat this growing menace.

We hypothesized that economic hardships drive fraud in society based on three categories of factors associated with identity fraud. First, we hypothesized that at least one of the macroeconomic factors – national unemployment rate, poverty rate, and median household income would be associated with identifying identity fraud. Second, we hypothesized that sociodemographic factors – education, online banking usage, financial debt carried by individuals, and the overall debt at a national level would be associated with explaining identity fraud. Lastly, we hypothesized that criminal behavior data – the number of fraud crime offenders, data breaches, and cyber-attacks would also be significantly associated with identity fraud.

We employed multiple linear regression methodology, with the number of reported identity fraud cases as the dependent variable, to ascertain the significance of the explanatory variables hypothesized to be associated with identity fraud. Our analysis revealed that the national unemployment rate,

https://doi.org/10.59720/23-154

online banking utilization, and the number of fraud crime offenders identified by law enforcement are statistically significant factors linked to identity fraud. This study highlights the complexity of comprehending identity fraud at a societal level, emphasizing the necessity to consider a broad spectrum of economic, social, behavioral, and criminal factors.

RESULTS

Results of Regression Analysis

To gain a statistically grounded understanding of the drivers of identity fraud, we based our analysis on data from 2005 through 2021 across the three broad categories of macroeconomic, socio-demographic, and criminal justice data. This choice was made to be expansive in terms of the range of explanatory variables we analyze with data collected over a significant period to ensure consistency of results over time. A total of 12 explanatory variables were chosen across these categories for our study. We used multiple linear regression methodology to identify explanatory variables that were both statistically significant (p-value ≤ 0.05) and were not correlated with each other (variance inflation factor (VIF) ≤10). The number of cases of identity fraud per 100,000 for each year was chosen as the predictor variable.

The regression results using all 12 variables indicated that the explanatory variables were correlated with each other and were not statistically significant in the presence of the other variables (**Table 1**). To model this complexity in results, we ran a backward elimination regression process to sequentially remove variables that were least significant based on their p-value and that had the highest correlation to each other based on their VIF values. After running the backward elimination process, we obtained a statistically significant model with an F-value of 41.297 and a p-value of < 0.001 (**Table 2**). The explanatory variables in the model were statistically significant and not correlated with each other. The explanatory variables that met these criteria were unemployment rate, percentage online banking, and number of financial fraud crime offenders (**Table 2**). We intentionally retained 'year' as a variable in our model even though it did not meet our significance and correlation thresholds. A fundamental assumption of linear regression is the independence of individual data points; however, in the context of identity fraud, this assumption may not hold true, as an individual could experience multiple instances of identity fraud over consecutive years. Consequently, the data points in the regression may lack mutual independence, violating this key assumption for multiple linear regression. To mitigate this issue, we included 'year' as a binary time variable. Given that our study aimed to identify significant variables in explaining identity fraud rather than numerically forecasting future volume of identity fraud, we deemed this approach acceptable for our analysis.

Variable category	Independent Variables	Unstandardized Coefficients		F-value	Significance p-value	VIF
		Coefficient	Standard Error			
Macro-economic	Unemployment rate	10.489	66.962	0.024649	0.883	11.796
	Poverty rate	104.638	259.815	0.162409	0.708	98.911
	Median household income	0.001	0.001	0.564001	0.494	90.17
Socio-Demographic	% Online banking	9.552	23.859	0.16	0.709	35.926
	% with college degree	18.976	273.936	0.004761	0.948	526.115
	Adult internet usage (18-29)	15.575	56.852	0.075076	0.798	191.848
	# credit cards per adult	0.455	5.993	0.005776	0.943	102.414
	Total US credit card debt	-0.004	0.039	0.010816	0.922	57.829
Criminal behavior	#Crime offenders	-0.002	0.002	1.382976	0.305	37.357
	# of cyber attacks	-5.37×10^{-05}	0	0.467856	0.531	17.574
	# of data breaches	-0.004	0.006	0.597529	0.483	48.029
Time	Year	-3.07	3.748	0.670761	0.459	18.395
	Constant	-39.575	113.282	0.121801	0.744	
Overall model				7.402	0.34	

Table 1: Key statistics for the initial run of the model with all twelve independent variables. The dependent variable measured was the number of reported identity fraud cases per 100k people. The p-values and VIF were all above the set thresholds for statistical significance and correlation. To model the complexity in the outputs, we followed this with a backward elimination regression procedure to isolate variables that were statistically significant and uncorrelated with other variables.

Variable category	Independent Variables	Unstandardized Coefficients		F-value	Significance p-value	VIF
		Coefficient	Standard Error			
Macro-economic	Unemployment rate	50.17	16.28	9.499	0.01	1.331
Socio-Demographic	% Online banking	7.916	3.495	5.13	0.043	1.472
Criminal behavior	#Crime offenders	-0.003	0	49.21	<.001	3.539
Time	Year	-0.373	1.258	0.088	0.772	3.957
	Constant	20.572	3.476	35.023	<.001	
Overall model				41.297	<0.001	

Table 2: Key statistics for final model after running the backward elimination procedure. The dependent variable measured was the number of reported identity fraud cases per 100k people. The p-values and VIF's are all within statistical significance and correlation thresholds. The only exception is the variable 'year' that we have kept on purpose to account for the time series nature of this dataset.

Macroeconomic factors

Our hypothesis was that macroeconomic factors like unemployment, poverty rate, and median household income significantly influence financial fraud. Of the three macroeconomic variables chosen for this analysis, only the national unemployment rate was statistically significant (p = 0.01, VIF = 1.331, **Table 2**). The two other variables, poverty rate and median household income, were heavily correlated with the unemployment rate (**Figure 1-2**). As part of the backward elimination process, these two variables had VIF > 10 and were removed from the final model. The other element of the result that we analyzed was the constant or coefficient for each variable of the final regression equation. The coefficient of the unemployment rate in the regression was positive at 50.77, indicating that higher unemployment rates are associated with higher levels of identity fraud.

Socio-demographic factors

We hypothesized that sociodemographic factors, namely education levels measured by the percentage with a college degree, online activity measured by the percentage of online banking and adult internet usage, and financial activity

JOURNAL OF
EMERGING INVESTIGATORS

https://doi.org/10.59720/23-154

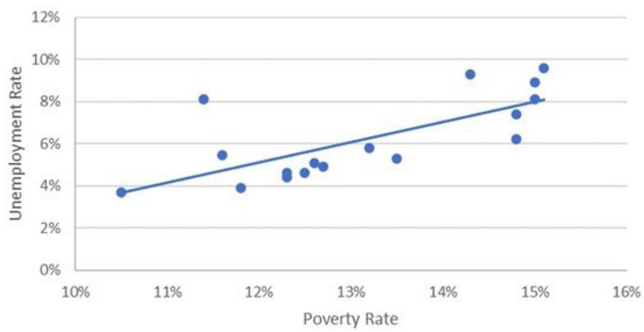

Figure 1: Correlation between the national unemployment rate and the national poverty rate. The coefficient of determination (R-squared value) is 0.5186 which means that 51.86% of the variation we observe in the unemployment rate is explained by the poverty rate (11, 12). Hence poverty rate was removed from the final model to reduce the impact of multicollinearity.

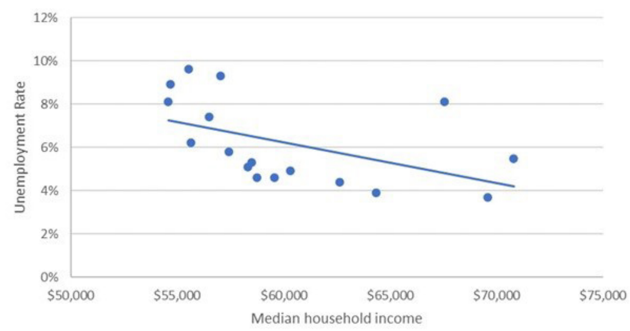

Figure 2: Strong negative correlation between the national unemployment rate and median household income. A strong negative correlation implies that having both variables in the model will distort an accurate interpretation of the model. The coefficient of determination (R-squared value) is 0.241 which means that 24.1% of the variation we observe in the unemployment rate is explained by the median household income (11, 12).

measured by the number of credit cards owned and total US national credit card debt, would significantly explain identity fraud. The only variable chosen based on significance and collinearity was the percentage of online banking ($p = 0.043$, VIF = 1.472, **Table 2**). The backward elimination process eliminated all the other variables in this category from the model due to the high correlation with online banking usage. The coefficient of percentage of online banking usage was positive at 7.916, indicating that higher online banking usage is associated with higher identity fraud.

Criminal behavior factors

We hypothesized that the effectiveness of the criminal justice system in our country, measured by the number of financial crime offenders, number of data breaches, and number of cyber-attacks, would significantly explain identity fraud. The only significant variable in this category was the number of financial crime offenders ($p < 0.001$, VIF = 3.539, **Table 2**). This variable also shows a strong correlation with our predictor variable, which was the number of identity fraud cases (**Figure 3**). Data breaches and cyber-attacks have become more common in the last few years, but these variables did not meet our statistical significance and correlation criteria and, hence, were not included in the final model. However, given these impacts on data security and privacy, their effect cannot be overlooked.

DISCUSSION

As we delve into the implications of our study's findings, it's pertinent to revisit the primary of our study. The primary purpose of our study was to understand the macro-level variables statistically linked to explaining identity fraud. The results of the study would inform future research and interventions at the individual and societal levels. We posited three main hypotheses.

Our first hypothesis was that macroeconomic indicators of financial hardship, specifically the unemployment rate, poverty rate, and median household income, would be significant in explaining identity fraud. Our analysis substantiates this hypothesis, revealing a correlation between identity fraud and financial hardship. The national unemployment rate emerged as statistically significant with a positive coefficient. In periods of elevated unemployment, individuals grappling with financial strain may resort to fraudulent activities to alleviate their circumstances. Conversely, heightened vulnerability during job searches may render individuals more susceptible to becoming victims of fraud, as they may inadvertently divulge personal information while seeking employment opportunities.

Our second hypothesis considered sociodemographic factors, specifically education level and financial activity. Education level was gauged by the percentage of the population with a college degree, online activity measured by the percentage of the population using online banking, and adult internet usage. Financial activity was assessed by the average number of credit cards owned by an individual and the total US credit card debt. Among these variables, only the percentage of online usage was statistically significant. This variable had a positive coefficient, implying that heightened online banking usage correlates with increased instances of identity fraud. A likely implication of our finding is that online banking adoption needs to occur with caution and greater security of our data. We rejected our hypothesis about education being a key driver of identity fraud due to the lack of statistical significance. A study conducted in 2021 revealed that targeted financial fraud education serves as an effective strategy for reducing vulnerability to fraud (6). However, for our study, data on financial fraud education was unavailable for the timeframe under examination. Nonetheless, the implications of our study underscore the significance of financial education as a potential deterrent in mitigating financial fraud.

Our third hypothesis was centered on criminal behavior. The only statistically significant variable in this category was the number of financial fraud offenders with a negative coefficient. This result suggests that as more perpetrators of financial fraud are apprehended, instances of identity fraud decrease. Interestingly, the number of data breaches and cyber-attacks were not statistically significant. A 2019 study by McKinsey and Company, drawing on data from 2018 to 2019, underscored the growing significance of data breaches as a driver of financial fraud, including identity fraud (7). However, our study's lack of statistical significance may be attributed to the relatively stable nature of data breaches during most of the examined period, with a notable surge observed

https://doi.org/10.59720/23-154

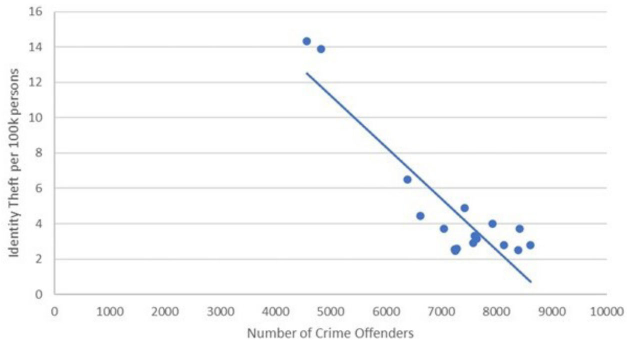

Figure 3: Strong negative correlation between identity theft per 100k persons and number of crime offenders. As more fraud criminals are apprehended, incidences of identity fraud decrease. The coefficient of determination (R-squared value) is 0.7987 which means that 79.87% of the variation in identity theft cases is explained by the number of crime offenders (3, 17).

only since 2017. Despite the statistical insignificance, the substantial impact of these breaches cannot be discounted; exposure to bad actors heightens vulnerability to identity fraud. Conversely, the significant role played by the number of criminal offenders emphasizes the need for a robust criminal justice system in combating identity fraud. An opportunity exists to bolster our criminal justice system to escalate the costs associated with perpetrating fraud, thereby deterring fraudsters more effectively.

It is important to acknowledge the limitations inherent in our study. The availability of accessible data constrained our findings; we could only construct our dataset dating back to 2005. Additionally, limitations existed regarding the types of variables obtainable from public sources. For instance, access to more comprehensive demographic data on both victims and perpetrators of identity fraud would have provided valuable insights. Regarding our modeling approach, we utilized a time-based indicator variable to accommodate the time-series nature of identity fraud. Given our focus was solely on analyzing the statistical determinants of identity fraud, this method suited our objectives. Moving forward, we aim to gather more extensive data and refine our methodology. Incorporating conventional time series analysis techniques will enable us to comprehend the drivers of identity fraud more comprehensively with more variables and make forecast predictions regarding future volumes of expected identity fraud. We expect a more granular analysis will facilitate more informed interventions at both the individual and societal levels.

In summary, our study analyzed the determinants of identity fraud in the US across a comprehensive range of variables over time. Research conducted in 2008 by Cebula and Koch highlighted the undocumented immigrants and state-level unemployment rate as the most influential predictors of identity fraud (8). Our study builds upon this research, offering a more expansive macro-level understanding of identity fraud's complexities. The pervasive nature of identity fraud erodes individuals' trust in governmental and institutional entities regarding data security and privacy (9). A 2017 study revealed Americans' inadequate knowledge of cybersecurity (9). Investments should be made by the government, education, and other large institutions to invest more in their cybersecurity and impart knowledge in this space.

MATERIALS AND METHODS

Data

The dependent variable chosen for our analysis was the reported number of identity fraud cases. These data are collected by the FTC through self-reporting of fraud from victims. At the individual consumer level, fraud is under-reported, primarily because victims often feel a sense of guilt and responsibility for the fraud (10). This self-reporting bias is a factor in this study but given the purpose of our study was to find variables statistically associated with identity fraud and not a numerical prediction of future fraud, we would be directionally accurate. To ease dealing with numbers in several hundred thousand, the number of reported identity fraud cases was converted to the number of reported identity fraud cases per 100,000.

We chose three broad categories of independent explanatory variables for this study: macroeconomic variables, socio-demographic variables, and criminal behavior. For the macroeconomic variables category, there were three variables based on the hypothesis that poverty and economic hardships drive fraud in society. These variables were the average national yearly unemployment rate, average national annual poverty rate, and median national household income (11, 12).

For the socio-demographic variables category, we chose five variables. These variables were the percentage of the population with a college degree, adult internet usage, the percentage of the population using online banking, the number of credit cards per person, and credit card debt (13-15). Adult internet usage was measured as the percentage of adults within the 18-29 age group that were internet users. It is worth noting that data from this source was missing for 2008 and 2009 (14). Given the general trend in the data, we interpolated data for those years to ensure the completeness of the dataset. Analysis done on the 18-29 age group showed a very strong positive correlation with other age groups. Rather than using each age group as a separate explanatory variable, we used the 18-29 age group as representative of adult internet usage to avoid the impact of multicollinearity in the analysis. For percentage of the population using online data, we had data missing for 2014. Similar to our approach for adult internet usage, we interpolated that data given the rising trend in the data.

For the criminal behavior category, we used three variables. The three variables we chose were the number of criminal fraud offenders, the number of data breaches, and the number of cyber-attacks reported by Federal agencies (16, 17).

Analysis methodology

Multiple linear regression was used as the core method for our analysis. The timeframe of this study was based on yearly data from 2005 through 2021 across the three categories. Based on the literature review, fraud victims can be revictimized for multiple years, hence there is a likely correlation between identity fraud data across multiple years (18). This violates the assumption for linear regression, which assumes that individual data points in the dependent variable are mutually independent. To account for the time-series nature of this problem, we created a binary indicator variable for years prior to and after 2018. The year 2018 was chosen because identity fraud cases went up sharply during that

https://doi.org/10.59720/23-154

period and beyond. This could be partly explained by better data collection methods employed by FTC and in part by a general increase in identity fraud. We used multiple linear regression with a backward elimination process to identify statistically significant independent variables ($\alpha = 0.05$) in predicting identity fraud. Given the nature of the variables we had chosen, e.g. unemployment rate and poverty rate, we had a hypothesis that some of these variables would be correlated with each other and may artificially inflate the fit of the model and hence we coupled this analysis with multi-collinearity analysis. We ran an ANOVA test with all the explanatory variables using Statistical Package for Social Sciences (SPSS) software. We used the backward elimination method to eliminate variables based on significance (F-value and p-value) and collinearity based on variance inflation factor (VIF 10). We iteratively did this exercise to get to a model with the best fit that, along with statistical significance, also made sense in terms of the sign of the coefficients. To account for the time-series impacts mentioned earlier, we chose to keep the binary period indicator variable irrespective of significance.

ACKNOWLEDGMENTS
The authors would like to thank Ms. Nirali Jantrania from Deep Run High School, Glen Allen, Virginia for reviewing the final paper and providing valuable insights.

Received: June 16, 2023
Accepted: December 30, 2023
Published: October 9, 2024

REFERENCES
1. Koops, Bert-Jaap, and Ronald Leenes. "Identity Theft, Identity Fraud and/or Identity-related Crime." Datenschutz Und Datensicherheit - DuD, vol. 30, no. 9, Sept. 2006, pp. 553–56. https://doi.org/10.1007/s11623-006-0141-2
2. US Bureau of Justice Statistics. "Victims of Identity Theft, 2021." Bureau of Justice Statistics, 12 Oct. 2023, bjs.ojp.gov/press-release/victims-identity-theft-2021. Accessed March 27, 2024.
3. "Consumer Sentinel Network Data Book 2022." Edited by Consumer Sentinel Network, *Consumer Sentinel Network*, Federal Trade Commission, Feb. 2023, www.ftc.gov/system/files/ftc_gov/pdf/CSN-Data-Book-2022.pdf. Accessed March 27, 2024.
4. IRTC. "ITRC Annual Data Breach Report - ITRC." *2023 Data Breach Report*, Identity Theft Center , Jan. 2023, www.idtheftcenter.org/publication/2023-data-breach-report/. Accessed March 28, 2024.
5. Deevy, Martha, et al. "Scams, Schemes & Swindles." *Scams, Schemes & Swindles: A Review Of Consumer Financial Fraud Research*, Financial Fraud Research Center, 2012,longevity.stanford.edu/wp-content/uploads/2017/01/Scams-Schemes-Swindles-FINAL-On-Website.pdf. Accessed March 28, 2024.
6. Burke, Jeremy, et al. "Can Educational Interventions Reduce Susceptibility to Financial Fraud?" *Journal of Economic Behavior & Organization*, vol. 198, 10 May 2021, pp. 250–266. https://doi.org/10.1016/j.jebo.2022.03.028.
7. Hasham, Salim, et al. "Financial Crime and Fraud in the Age of Cybersecurity." *McKinsey & Company*, 1 Oct. 2019, www.mckinsey.com/capabilities/risk-and-resilience/our-insights/financial-crime-and-fraud-in-the-age-of-cybersecurity. Accessed March 29, 2024.
8. Cebula, Richard J., and James V. Koch. "The Impact of Undocumented Immigration on ID Theft in the United States: An Empirical Study." *The Open Economics Journal*, vol. 1, no. 1, 25 Sept. 2008, pp. 37–46. https://doi.org/10.2174/1874919400801010037.
9. Olmstead, Kenneth. "What the Public Knows about Cybersecurity." *Pew Research Center: Internet, Science & Tech*, Pew Research Center, 22 Mar. 2017, www.pewresearch.org/internet/2017/03/22/what-the-public-knows-about-cybersecurity/. Accessed March 29, 2024.
10. Schoepfer, Andrea, and Nicole Leeper Piquero. "Studying the Correlates of Fraud Victimization and Reporting." *Journal of Criminal Justice*, vol. 37, no. 2, 14 Mar. 2009, pp. 209–215. https://doi.org/10.1016/j.jcrimjus.2009.02.003.
11. "Unemployment rate - 14 years and over." *Labor Force Statistics from the Current Population Survey*, United States Bureau of Labor Statistics, 2023. data.bls.gov/timeseries/LNU04023554&series_id=LNU04000000&series_id=LNU03023554&series_id=LNU03000000&years_option=all_years&periods_option=specific_periods&periods=Annual+Data. Accessed March 29, 2024.
12. United States Census Bureau, United States Department of Commerce, et al. *Income in the United States: 2022*, United States Census Bureau, Sept. 2023. www.census.gov/content/dam/Census/library/publications/2023/demo/p60-279.pdf. Accessed April 1, 2024.
13. "Percent of People 25 Years and Over Who Have Completed High School or College, by Race, Hispanic Origin and Sex." *CPS Historical Time Tables*, United States Census Bureau, 2023. www.census.gov/data/tables/time-series/demo/educational-attainment/cps-historical-time-series.html. Accessed April 1, 2024.
14. Fox, Susannah. "51% of U.S. Adults Bank Online." *Pew Research Center: Internet, Science & Tech*, Pew Research Center, 7 Aug. 2013, www.pewresearch.org/internet/2013/08/07/51-of-u-s-adults-bank-online/. Accessed April 1, 2024.
15. "Household Debt and Credit Report." *Household Debt and Credit Report - FEDERAL RESERVE BANK of NEW YORK*, Federal Reserve Bank of New York, 2023, www.newyorkfed.org/microeconomics/hhdc. Accessed April 2, 2024.
16. Schmitt, Glenn R., and Amanda Russell. "Fiscal Year 2020 - Overview of Federal Criminal Cases." *Overview of Federal Criminal Cases*, United States Sentencing Commission, Apr. 2021, www.ussc.gov/sites/default/files/pdf/research-and-publications/research-publications/2021/FY20_Overview_Federal_Criminal_Cases.pdf. Accessed April 2, 2024.
17. The United States Office of Management and Budget (OMB), The United States White House. *FEDERAL INFORMATION SECURITY MODERNIZATION ACT OF 2014 ANNUAL REPORT FISCAL YEAR 2022*, The Office of Management and Budget, 2023. www.whitehouse.gov/wp-content/uploads/2023/05/FY22-FISMA-Report.pdf. Accessed April 4, 2024.
18. Button, Mark, et al. "Fraud Typologies and the Victims

https://doi.org/10.59720/23-154

of Fraud: Literature Review." *University of Portsmouth*, National Fraud Authority, 8 Dec. 2014, researchportal.port.ac.uk/en/publications/fraud-typologies-and-the-victims-of-fraud-literature-review. Accessed April 4, 2024.