The University of Texas at Austin
Center for Identity

Personal Data Early Warning System: Machine Learning Models Extract Identity Theft and Fraud Trends from News

Razieh Nokhbeh Zaeem
K. Suzanne Barber
Jessica Cruz-Nagoski
Luke Norrell
Michael Sullivan
Jonathan Walsh
Dylan Wolford
Yasira Younus

UTCID Report #21-04

August 2021

Personal Data Early Warning System: Machine Learning Models Extract Identity Theft and Fraud Trends from News

Razieh Nokhbeh Zaeem
nokhbeh@utexas.edu
Center for Identity
The University of Texas at Austin
Austin, Texas, USA

Jessica Cruz-Nagoski[*]
Luke Norrell[*]
Michael Sullivan[*]
Jonathan Walsh[*]
Dylan Wolford[*]
Yasira Younus[*]
jessicacruzn@utexas.edu
lukenorrell@utexas.edu
sullivan.michael57@gmail.com
jonathan.walsh@utexas.edu
dylanwolford@utexas.edu
yasirayounus@utexas.edu
The University of Texas at Austin
Austin, Texas, USA

K. Suzanne Barber
sbarber@identity.utexas.edu
Center for Identity
The University of Texas at Austin
Austin, Texas, USA

ABSTRACT

Each year, cyber attacks pose a greater and greater risk to consumer personal information stored by corporations and government agencies. Billions of consumer records are breached each year and data breaches compromise the personal data of hundreds of millions of citizens. These breaches are extremely costly–financially and in terms of privacy and reputation–to people (through identity theft and fraud) and to companies (through the abuse of their collected information for which they are accountable). What is more, the theft of data often acts as a gateway in the complex and interdependent ecosystem of personal data. Personally Identifiable Information (PII) is breached to gain access and steal more PII in a chain of events and tactics. Therefore, there is a need to build tools to help people and businesses navigate the dangerous waters of identity theft and fraud. The cyber world, however, is an evolving landscape and trends change often. People and organizations need to have a current and accurate situational awareness understanding trends such as common breach threats and tactics, types of data most frequently attacked, and personal information most often exposed with the highest negative consequences.

Enter the Personal Data Early Warning System (PDEWS), an online dashboard that tracks and displays the current cyber threat landscape and generates actionable insight into trends and patterns. PDEWS exists as an automated pipeline, collecting data each day about ongoing cyber threats. There are four major phases of PDEWS. First, PDEWS prowls through daily identity theft and fraud news stories and scrapes the body text. Then it formats the text into the representation required for a machine learning application and places that text in an Amazon Web Services cloud infrastructure. Next, PDEWS applies machine learning models trained on a private identity theft article corpus to extract relevant threat labels. Finally, PDEWS displays those trends on an online dashboard alongside recommendations researched to have the greatest mitigation capabilities against the current threat landscape. PDEWS specifically highlights PII used to gain access, PII stolen as a result, and the steps and tactics used in between, to shed light on the interdependent nature of identity theft and fraud. Publicly available at https://pdews.herokuapp.com, the PDEWS system stands as a novel approach to analyzing cyber threat trends via online news while delivering threat mitigation recommendations based on best practices.

CCS CONCEPTS

• **Information systems** → *Web searching and information discovery*; *Data extraction and integration*; • **Security and privacy** → *Economics of security and privacy*; *Privacy protections*; • **Applied computing** → Surveillance mechanisms; • **Social and professional topics** → **Identity theft**.

KEYWORDS

machine learning, identity theft and fraud, news articles, interdependent privacy, data science

1 INTRODUCTION

Identity theft, fraud, abuse and exposure (hereafter identity theft for short) involve the use of a victim's identity, particularly the use of the victim's Personally Identifiable Information (PII), without permission. Such information may get compromised in a variety of ways: e.g., through social engineering methods or the breach of consumer information databases. Once compromised, a person's PII can be misused in many ways: e.g., an identity thief might put fraudulent charges on existing accounts, open new accounts in the victim's name, or file a tax return and collect a refund. Identity theft crimes harm victims in ways that include, but are not limited to, invasion of privacy, financial loss, loss of physical or intellectual property, reputation damage, effort and money spent to recover from the incident and prevent further misuse of the compromised PII, and emotional distress [27].

[*] These authors contributed equally to this research.

Unfortunately, identity theft and fraud crimes are increasingly common. According to the 2016 U.S. National Crime Victimization Survey [36], at least 25.9 million Americans were affected by identity theft and fraud in the previous year. In the consumer sentinel network from the Federal Trade Commission (FTC), identity theft and fraud are one of the top categories of scam reported to the agency, to which Americans lost more than $1.9 billions in 2019. Identity theft is a costly problem with constantly evolving patterns of criminal tactics and behaviors [37].

As more businesses and people become victims of identity crimes, it is increasingly important for businesses and people to better understand the crimes of identity theft, fraud, and abuse. While statistics have been collected regarding the number of exposed records or the financial loss to individuals [21, 30], we are not aware of any *publicly available* system that analyzes *up-to-date* identity theft data and *fine tunes statistics on demand.* **We present PDEWS (Personal Data Early Warning System), our publicly available[1] online dashboard that automatically collects the semi-live stream of online news articles about identity theft and fraud, analyzes these news articles with machine learning, and displays fine tuned statistical trends (e.g., in a particular market sector).** PDEWS particularly drills down into the interdependent nature of PII: It analyzes PII used to gain access, PII subsequently exposed and stolen, the steps the criminals took in between, and the tactics they used.

We organize the rest of this paper as follows. Section 2 briefly reviews related work on methods of data collection for identity theft and fraud, as well as the application of machine learning in their analysis. Section 3 covers necessary background on our previous corpus of identity theft news articles, used to train machine learning models here. Section 4 covers the design and implementation of PDEWS, and Section 5 concludes the paper and highlights some future work directions.

2 RELATED WORK

In this section, we summarize related work on methods previously used to collect identity theft and fraud data, as well as the application of machine learning on the collected data.

2.1 Collection of Identity Theft and Fraud Data

Several years ago, various authors pointed out that few details were known about the actual methods used by perpetrators of identity theft and fraud [19, 29], and this is still largely true today. To know more about identity theft and fraud, previous work has introduced several avenues of collecting identity theft data.

2.1.1 Agency Data. The first commonly used source of identity theft data is gathered from agencies [1, 15, 18]. For example, the FTC's [15] database gathers identity theft complaints from FTC's telephone- and web-based complaint systems in addition to more than one hundred federal and state organizations. Even though some have raised concerns [29] towards the representativeness of Consumer Sentinel Network and other agency data (including lack of consistency in the definition of identity theft, under-reporting identity theft both from consumers and agencies, and bias due to change in consumer awareness and agency policies) researchers still use agency data widely, while making an effort to manage the amount of data in such databases [31, 32].

2.1.2 Surveys. Synovate on behalf of the FTC, [23], and several other universities and research organizations have conducted national surveys about identity theft and fraud. Apart from great variance in their sample sizes and some differences in methodologies, such surveys have been criticized [29] for issues such as non-response bias, difficulty to contact victims (especially because victims of identity theft sometimes have to change their contact information), and relying solely on the memories of victims.

2.1.3 Interviews. While interviews with *victims* is another means of collecting identity theft data for research in academia [2, 5, 26], there is some research that is based on interviews with identity *thieves and fraudsters* [16]. Such interviews can provide great details about the methods these criminals use. Still, this type of research is likely to be limited by small sample sizes and skewed by the fact that their subjects are all perpetrators who were caught, incarcerated, and willing to be interviewed.

2.1.4 Reports from Affected Organizations. Organizations affected by data breaches that lead into identity theft and fraud sometimes report their own data (e.g., [3, 17]).

2.1.5 News Stories. News stories have been used for identity theft data collection [28, 41, 42][2]. News stories have several characteristics that make them an appropriate source of data: There is a tremendous amount of news stories about identity theft; They are widely available, as opposed to agency data that is difficult to obtain from government agencies or corporations; Finally, most news stories are reliable and trustworthy since the news media is responsible for providing accurate information to the public and are held accountable. This source complements other sources by focusing on news articles that narrate a wide range of identity theft stories, from victims, law enforcement, and companies. This source, too, has some bias. The news media tends to report stories that are considered newsworthy. Also, the same news story might get reported in several forms.

We base PDEWS on news stories as the source of our data and to train our machine learning models (Section 3). PDEWS, however, is not inherently limited to news stories. One could apply our machine learning techniques on other data sources of identity theft and fraud or supplement the current data source of news articles with new sources.

2.2 Application of Machine Learning in The Analysis of Identity Theft and Fraud Data

Irrespective of the data source, researchers have leveraged data mining and machine learning in order to prevent and combat identity theft in various ways. Examples include:

- Facilitating the tasks performed by law enforcement, e.g. Chen et al. [14] built a general framework to help Arizona police departments investigate a wide range of crimes including but not limited to identity theft.

[1] https://pdews.herokuapp.com

[2] Our previous work [41] which appeared in Computers & Security in 2017 pioneered the use of text mining on online news stories that report on the topic of identity theft.

- Detecting phishing attacks [4, 22] and identity theft that occurs through it.
- Detecting credit card fraud, a subject closely related to identity theft, e.g., through using the data mining approaches support vector machines and random forests [6].

In particular, few researchers have previously used machine learning to model and analyze *identity theft stories* [7, 20, 41]. Their work, however, was performed in an off-line manner that did not produce actionable, fine-tuned, and live trends of identity theft to the public.

3 BACKGROUND: CORPUS OF IDENTITY THEFT STORIES TO TRAIN MACHINE LEARNING MODELS

We utilize our prior work on Identity Threat Assessment and Prediction (ITAP) to train machine learning models. ITAP [39, 40, 42] in an ongoing project that collected and modeled over 6,000 identity theft news stories that occurred over the past twenty years (from 2000 to 2020). These news articles were discovered through RSS (Rich Site Summary) feeds from the Internet. A team of modelers read each of these 6K+ news articles over a period of six years (from 2014 to 2020). The modelers have manually extracted over 50 details about each identity theft incident reported in the news. These details include:

- the type of the incident
- how and when the incident happened
- the methods and resources used to carry out the crimes
- the vulnerabilities exploited
- the types of personal information compromised
- the demographics of the victims
- the consequences for the victims and perpetrators

ITAP has been used in various applications [8–11, 13, 24, 25, 33, 34, 38]. In this work we take advantage of its rather large body of manually labeled identity theft incidents for training, cross validation, and testing machine learning models.

4 PDEWS

Personal Data Early Warning System (PDEWS) features a pipeline system with four steps:

(1) finding identity theft news stories on the Internet,
(2) extracting the news story's body text,
(3) running trained machine learning models on the extracted text to generate data labels,
(4) displaying relevant information on a dynamically updated, publicly available dashboard.

To accomplish the end goal of a daily-updating dashboard, we used a pipeline approach to set up the infrastructure to scrape news articles and run trained models on the article text daily. Before these steps, we trained five classification models on the ITAP data to predict certain desired labels. Then, (1) our pipeline runs once per day and scrapes RSS feeds that are tuned to keywords relating to identity theft. (2) We extract and pre-process the body text from the articles. (3) The trained models are then run on the preprocessed data and append new stories and model outputs to the database along with the date the article was scraped. (4) The dashboard then can access this final database with all of the stories, body text, and extracted labels to aggregate and display for the end user. This pipeline model allows us to graph and display various beneficial insights that will be of use to our audience (including people and businesses), such as identifying increasingly vulnerable PII in a specific sector over the past month or quarter, or examining an industry's recent track record with cyber security events.

4.1 Finding Identity Theft News

The beginning of the pipeline is the data gathering stage where new news articles are gathered once per day to get a daily view of new emerging threats via news stories. We set up three RSS feeds that search for the words "identity theft", "identity thieves", and "identity fraud" and return links to news articles about those topics. We ensure that we eliminate obviously duplicate news articles.

4.2 Extracting News Body Text

From the news article links, we used the *Python requests* library to get the raw HTML of the news story. Within the raw HTML, there is a lot more information than just the text of an article, such as the font, font size, and page layout. One of these attributes is the response code, which indicates whether or not a request was successful, where a response code in the 400's or 500's indicates failure. After filtering out the unsuccessful responses, we extracted the body text with the *newspaper3k Python* package. This package is specifically built to extract aspects from news articles including body text. The body text for each article, along with the extraction time-stamp, is then saved to a CSV file on Amazon Web Services (AWS) S3 for future processing steps.

4.3 Running Machine Learning Models

Once we have the text of the identity theft news story, we run trained models to generate data labels for each story.

4.3.1 Model Training. We trained five models that are predicting five different types of labels from our news articles. The labels our models are predicting from each news story are (1) the PII used to gain access, (2) PII stolen after gaining access, (3) tactics used by the attacker, (4) industry sector the attack occurred in, and (5) the steps the attackers took. Each label has a set number of values that the model can predict. These label and their values come from the manual labeling of ITAP and are listed in Table 1.

All of the models are multi-label classification models except the sector model, which is a multi-class classification model. For the multi-label classification models, We used a One-vs-Rest Classifier from *sklearn* with a logistic regression estimator. For the multi-class model to predict which sector the story was in, a Gradient Boosting Classifier was used.

We utilized the ITAP data to train the models. We created three TF-IDF vectorizers for our models after cleaning up and formatting the original ITAP dataset to suit the needs of each model. We preprocessed all the sample news stories with the following steps: remove everything that is not an alphanumeric character, remove white spaces, convert text to lowercase, lemmatize using *WordNetLemmatizer()*, and remove common English stop words. Once the preprocessing was completed, we applied our vectorizing

Table 1: Possible labels to predict for each model, based on ITAP training data.

Industry Sector	PII Used to Gain Access	Tactics Attackers Used	PII Stolen	Steps	Steps (cont.)
Healthcare and Public Health	Academic Info	Access Misuse	*Same as PII Used*	Abuse	Misplace
Consumer/Citizen	Account Access Info	Audio/Visual Involvement		Acquire	Monitor
Government Facilities	Bank Access Info	Broken Into		Act to Elicit Response	Neglect
Education	Credit Info	DDOS		Act upon	Purchase
Commercial Facilities	Criminal Info	Device Mishandled		Activate	Record
Financial Services	Customer Info	Email Scam		Alter	Recruit
Information Technology	Diagnosis Data	Impersonation		Analyze	Request
	Employee Info	Malicious Link		Block	Scenario
	Fraudulent Info	Malware		Breach	Sell
	General Account Info	Misinformation		Break into	Send
	General Bank Account Info	Other		Communicate	Steal
	General Business Info	PII/Credential Stolen		Compile	Submit
	General Insurance Info	Phishing		Conceal	Surveil
	General Personal Info	Phone Call Scam		Coordinate	Transfer
	General Personal Medical Info	Ransomware		Create	Upload
	Loan Data	Removable Media		Deactivate	
	Medical Insurance Info	Security Vulnerability/Mismanage		Decide	
	Medical Test Results	Social Media Involvement		Destroy	
	Medication Info	Synthetic Info		Disable	
	Miscellaneous Money Info	Transfer		Discover	
	Official Identification Data			Expose	
	Other			Find	
	Personal Life Info			Function	
	Personal Location Info			Impersonate	
	Personal Phone Data			Infect	
	Relative's Personal Info			Inflict Punitive Measure	
	Specific Medical Service Data			Leak	
	Travel Info			Lie	
	Vehicle Info			Malfunction	
	Tax Info			Mismanage	

Table 2: Classification results (weighted averages) for the machine learning models.

Model	Precision	Recall	F1	Accuracy
Industry Sector Multi-Class	0.587	0.590	0.576	0.590
PII Used to Gain Access Multi-Label	0.335	0.151	0.185	0.831
Tactics Attackers Used Multi-Label	0.732	0.253	0.345	0.855
PII Stolen Multi-Label	0.654	0.339	0.396	0.849
Steps Multi-Label	0.718	0.514	0.545	0.919

function to the story text and proceeded to train our models according to the specifications above. The vectorizers were later used to transform our new dataset before our models made predictions on them.

To evaluate our models, we focused on using the accuracy, precision, recall, and F1 scores for all the models and the hamming distance and hamming loss score for just the multi-label classification models. For our one multi-class model that is predicting the sector for each story, we did an 80/20 split on the dataset for training and testing data, then we found the F1 score by comparing the predicted values with the actual values. Table 2 shows the scores for the Industry Sector model.

To evaluate the four multi-label classification models, we created a pipeline that predicted each label one at a time on our pre-split dataset, then we measured the accuracy, precision, recall, and F1 scores for each label. Finally, we calculated the hamming loss and hamming distance score for the entire model with all the labels predicted. The hamming loss looks at the fraction of labels that were incorrectly predicted, and the closer to zero the score is, the more accurate the model is. The hamming distance is another scoring metric popular among multi-label models—the closer the hamming distance is to one, the more accurate the model. Investigating the scores calculated above, we found that the models did not perform well, in part because our ITAP training data is sparse: we saw that the labels that had the most training data performed the best.

To compensate for training data sparsity, for the multi-label classification models, we oversampled using the Multi-Label Synthetic Minority Oversampling Technique (MLSMOTE) [12] technique for the labels that did not have enough instances in the preprocessed ITAP dataset, in order to provide 1,000 more dataset points for our models. MLSMOTE is a quite popular oversampling technique, specifically for multi-label classification, and is an extension of the long time favorite SMOTE, an oversampling technique that is used for a wide variety of problems. MLSMOTE seeks to calculate the imbalance ratio between the numerous labels provided for the model at hand. Each data label is assessed individually, and then the average of the overall category is taken in order to make the ratio assessment. The ratio numbers are then used to calculate how much synthetic data to create to reduce the imbalanced data. We realized that, before oversampling, the multi-label classification models were only predicting well on the labels that had a lot of

Table 3: The improvement in the PII Used to Gain Access Model with MLSMOTE: Changes In Hamming Loss and Distance Scores

	Hamming Loss	Hamming Distance
Before MLSMOTE	0.059	0.121
After MLSMOTE	0.039	0.453

examples in the dataset. So we oversampled the data to improve classification.

Overall, even after oversampling, we saw that the labels that had the most training data to begin with performed the best. Since we knew that our ITAP dataset was indeed sparse, we had expected this result. For example, Table 3 displays the changes in the results in the PII Used to Gain Access model before and after oversampling. Our attempts to improve the models by oversampling did pay off, but more training data will be needed in the future in order to improve the models, especially for the labels with very small representations. Table 2 shows weighted averages for the multi-label models after oversampling. Note that there are many labels to predict. For example, PII used to gain access and PII stolen each could have 30 labels, tactics could have 20 labels, and steps could have 45 labels (Table 1). A random classifier has very low chances of predicting correct labels with these many possible values.

4.3.2 Model Execution. The data processing stage of the pipeline uses each day's new articles and the trained models to extract desired labels from each article. The text of the article must first be converted into the correct input format for the machine learning models. Then, each of the five models is run on the preprocessed data to label each article with the desired final labels. The final extracted labels are then published to a database where the dashboard back-end can access the labeled data. This stage along with the data gathering stage are all implemented on AWS using AWS Step Functions, AWS Lambda, and S3.

Preprocessing of the text is largely standardized for all the models, utilizing standard NLP preprocessing techniques, with the exception of some model-specific vectorizers. For each story, we remove non-alphanumeric symbols, remove stopwords, and then lemmatize each word. This turns a sentence like "Seven hundred people's identities were stolen." into just the sequence of words "seven hundred people identity stolen". To get these sequences of words into a format for machine learning, we have three pre-fitted TF-IDF vectorizers stored on an AWS S3 bucket. We load each of these vectorizers and use them to transform each sequence of words in the machine learning specific format the models expect. Once story text for each article has been converted to the format expected by each model, we load all five of the models from an AWS S3 bucket and run the models on each story to generate the desired labels.

The actual AWS infrastructure consists of two AWS Lambda functions that run one after the other through a basic AWS Step Function, as shown in Figure 1. The first Lambda function implements the Data Gathering process. After completion of that Lambda Function, the Step Function triggers the second Lambda which pre-processes the text and runs the models on the new data. The extracted labels are uploaded to the data storage S3 bucket, accessible by the website back-end.

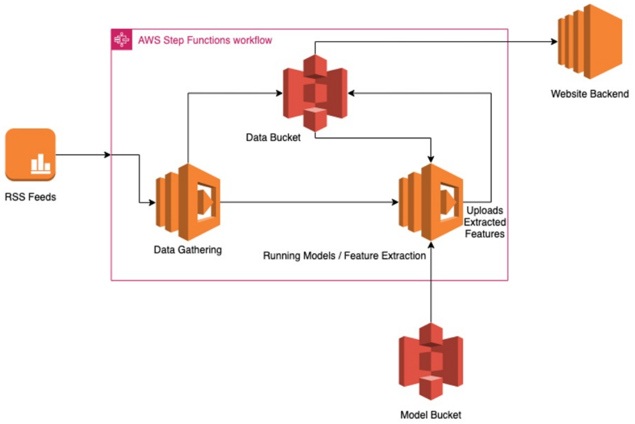

AWS Step Functions workflow
Website Backend
Data Bucket
RSS Feeds
Data Gathering
Running Models / Feature Extraction
Uploads Extracted Features
Model Bucket

Figure 1: Data gathering and feature extraction pipeline architecture.

4.4 Displaying Results

PDEWS incorporates a user interactive dashboard, which serves as a quick way to visualize all the threat information our models are collecting. Any changes in trends are now discovered quickly and easily. We broke up the development of the dashboard into two sections: the front-end and back-end. Both are hosted in separate areas, but connect with each other through numerous API calls and responses. Modularizing the development of the dashboard allows future developers to quickly modify the user interface or trends without having to change a significant portion of either component.

4.4.1 Dashboard Back-End. The dashboard back-end component of PDEWS serves as the hub of data calculation and communication across the pipeline. This component of PDEWS web application is written in *Python Flask*, and hosted on a Heroku server. It collects our newly labeled data from AWS cloud storage and performs all necessary calculations in order to create identity theft and fraud trends. Here, we can sort and filter data, and execute any data groupings or algorithms without affecting the performance of PDEWS user interface. This also ensures that the user has a smooth experience in loading all the components of the dashboard. Since we have to perform many API calls at once and handle large amounts of data, it is useful to do expensive functions on the server-side back-end.

The main purpose of our PDEWS dashboard is to display appropriate and actionable trends for users. We have a total of 14 charts with data that are updated daily, with calculation to draw these charts happening on the back-end. We have an endpoint for each of these graphs in the PDEWS back-end, which are used to create all the data to display. We also adopt a computationally intensive algorithm [35] that performs the calculations for our recommendation system portion of the dashboard. All of the endpoints also ensure that the data sent back is ready to display and formatted appropriately so that any components can load quickly.

4.4.2 Dashboard Front-End. The dashboard front-end is the user-facing final application deliverable of PDEWS. We used *React JS* to develop the front-end, and the server for it is hosted in Heroku as well. Using React ensures that we have components that load quickly and are built to be easily responsive in order to improve web application performance.

Our UI is divided into three pages. We have our home page (Figure 2), where we give a description of the project along with the motivations behind creating it. We also display two pie charts with real-time data from our back-end in order to give a quick introduction to the information we are displaying. Our home page is also where we integrated the recommendation system for our current data trends based on the work of Tyagi et al. [35]. We placed this system on the homepage since our dashboard is a warning and alerting system, and we want users to know important and actionable information as quickly as possible.

The bulk of our data resides in the threat trends page. Here, we display 12 total graphs divided into six rows, with two graphs on each row. This format is for comparison purposes: each row contains one graph that answers a threat question for a specific sector, and a second graph that answers the same question for all sectors. This way, the user can compare current threat trends from a specific sector to the overall threat landscape, and refer to the information from the sector in which they are interested. We address six types of questions through the graphs on the threat trends page:

(1) How many incidents have occurred? (Figure 3 answers this question for the Commercial Facilities Sector as an example. On the left the number of identity theft and fraud incidents in the selected sector are charted, accumulated in weeks. On the right the same info is shown for all the sectors combined.)
(2) What tactics are used to expose data? (Figure 4 answers this question for the Commercial Facilities Sector on the left and for all sectors on the right.)
(3) What kind of information was used to gain access?
(4) What are the trends of information used to gain access? (Same data as question 3, but with time as the x axis)
(5) What kind of information has been exposed?
(6) What are the trends of information that has been exposed? (Same as question 5, but with time as the x axis. Figure 5 answers this question for the Commercial Facilities Sector as an example. Questions 3, 4, and 5 have similar styles of graphs.)

By comparing these graphs between particular sectors and overall trends, we hope the user gains a better understanding of the threats they may encounter. There is a convenient drop down selector at the top of the threat trends page that allows the user to pick a sector, and that selection automatically modifies all the sector specific graphs to display the correct information. All the graphs have tool-tips that give further details about the information being displayed, e.g., how to read and interact with the chart, and what sort of information it is displaying. The final page of the PDEWS application is our about page.

Four of the five models explained in Section 4.3.1 were used to answer the above questions. (Question 1 uses the industry sector model, Question 2 uses the tactics model, Questions 3 and 4 use the model to find PII used to gain access, and finally Questions 5 and 6 use the model to detect PII stolen after gaining access.) The last model, the steps the attackers took, is instrumental to our recommendation system component on the home page.

For our recommendation system, we take advantage of previous work on I-WARN [35]. I-WARN is a set of algorithms capable of mapping open-source threat information to the MITRE ATT&CK framework. ATT&CK is a framework that helps understand lateral movement of an attack to offer mitigation and risk reduction tactics. Our recommendation systems takes the threat trends we inferred from news articles, applies the algorithms developed in I-WARN, and delivers the top three recommendations from ATT&CK to help mitigate future incidents. Each of these recommendations links to an entry in the widely used ATT&CK web page[3] of the MITRE non-profit organization. MITRE has implemented the ATT&CK Matrix, suggesting lists of possible horizontal movement throughout the incident response for any given cyber incident. It should be noted that the United States Cybersecurity and Infrastructure Security Agency (CISA) alerts currently apply the MITRE mitigation and detection techniques, indicating the matrix is being actively used in private and public sectors as guidelines for incident responses.

5 CONCLUSIONS AND FUTURE WORK

The Personal Data Early Warning System (PDEWS) provides actionable insight into identity theft trends. Our fully automated data gathering dynamically finds and pre-processes identity theft and fraud news articles online. Our system then stores these news articles where labels can be extracted seamlessly with machine learning. Our data processing component is an AWS infrastructure capable of housing and applying machine learning models to this corpus and importantly automates the entire process.

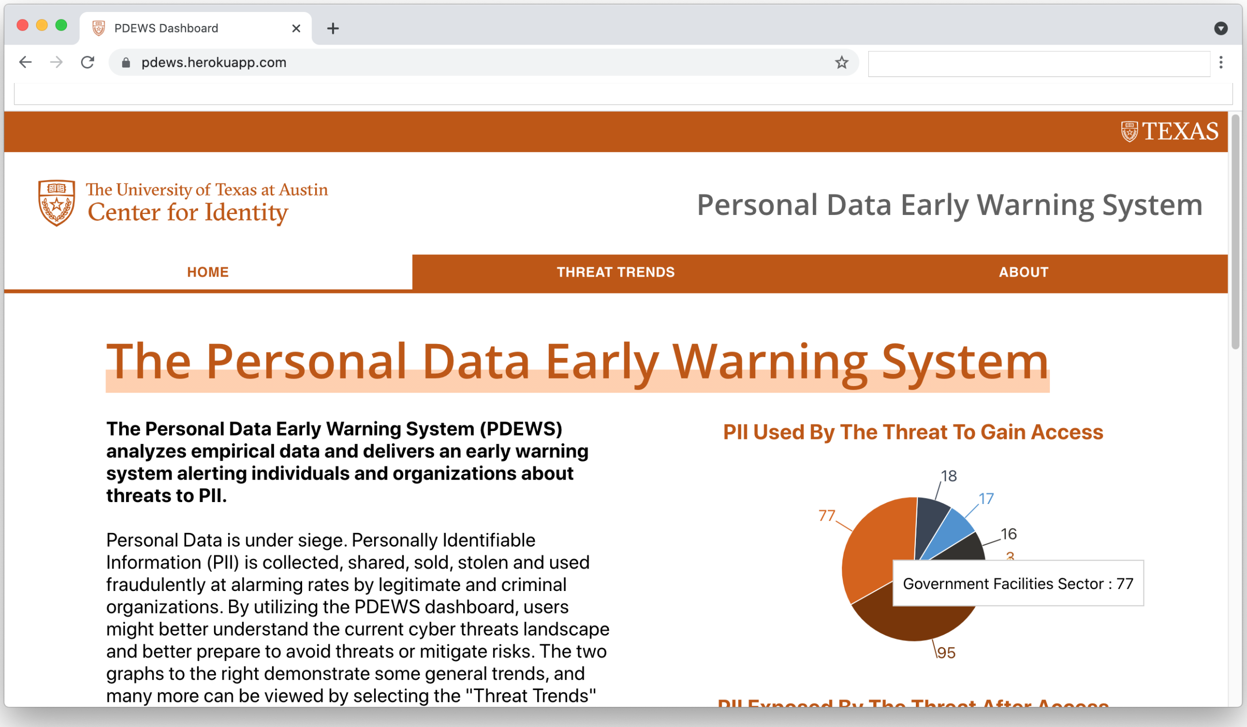
We trained multiple machine learning models on ITAP—a longitudinal dataset of over 6,000 identity theft and fraud news stories from 2000 to 2020, manually labeled by modelers. We deploy these models in our AWS online infrastructure for application on new articles. PDEWS collects related RSS feed of identity theft and fraud news stories every day and applies these models to the new articles and stores resulting labels in a dynamically updated format. The publicly available front-end of PDEWS automatically charts thread trends and displays recommended mitigation activities.

5.1 Limitations and Future Work

We envision multiple ways to improve PDEWS:

(1) As we discussed in Section 2.1, news articles are not the only source of identity theft and fraud data, nor are they the most comprehensive source. Other sources can complement these news articles. Fortunately, PDEWS can work with any other source(s) of identity theft and fraud data as long as the data is fed into its machine learning models for training, trend extraction, or both. Using other sources of data is a promising future work direction.
(2) We plan to add additional RSS feeds that look for different keywords; all of the feeds we have set up search for words similar to identity theft but not more varied cyber security related keywords. Expanding the RSS feeds to other RSS

[3] https://attack.mitre.org


PDEWS Dashboard
pdews.herokuapp.com
TEXAS
The University of Texas at Austin
Center for Identity
Personal Data Early Warning System
HOME
THREAT TRENDS
ABOUT
PII Used By The Threat To Gain Access
18
17
16
3
77
95
Government Facilities Sector : 77
PII Exposed By The Threat After Access

The Personal Data Early Warning System

The Personal Data Early Warning System (PDEWS) analyzes empirical data and delivers an early warning system alerting individuals and organizations about threats to PII.

Personal Data is under siege. Personally Identifiable Information (PII) is collected, shared, sold, stolen and used fraudulently at alarming rates by legitimate and criminal organizations. By utilizing the PDEWS dashboard, users might better understand the current cyber threats landscape and better prepare to avoid threats or mitigate risks. The two graphs to the right demonstrate some general trends, and many more can be viewed by selecting the "Threat Trends"

Figure 2: A screen-shot of the PDEWS front-end: the Home page.

feeds or additional sources would make the overall database of news stories more robust and accurate to the threat landscape.

(3) Even tough our training dataset of ITAP contained thousands of manually labeled news stories, it was sparse for some labels. Due to the limitations of our training dataset, our model accuracy was constrained. In the future, a better training dataset will be beneficial in order to improve the models. Since such a dataset does not currently exist publicly, one will have to be created. If such a dataset is created with the objectives of PDEWS in mind, the model accuracies will improve significantly. Our main suggestion is to label appropriate news stories by manually identifying exact words and phrases in a body of text that result in a specific label. With this type of labeling, the model will be able to predict labels with much higher accuracy than if given only label categories and a large body of text. Once the dataset is created, a Named Entity Recognition model can be used to quickly identify appropriate labels. With this technique, it will more than likely not be required to label thousands of articles in order to achieve reasonable precision and recall.

Our hope is that the insights extracted by PDEWS can facilitate a better response by users, such as information security professionals, to emerging cyber threats and can help mitigate risks. Finally, our work sheds light on the interdependent ecosystem of private information, particularly considering the fact that identity theft actors obtain new PII based on previously exposed personal information.

ACKNOWLEDGMENTS

This work was in part funded by the Center for Identity's Strategic Partners. The complete list of Partners can be found at https://identity.utexas.edu/strategic-partners.

REFERENCES

[1] Stuart FH Allison, Amie M Schuck, and Kim Michelle Lersch. 2005. Exploring the crime of identity theft: Prevalence, clearance rates, and victim/offender characteristics. *Journal of Criminal Justice* 33, 1 (2005), 19–29.

[2] Keith B Anderson, Erik Durbin, and Michael A Salinger. 2008. Identity theft. *Journal of Economic Perspectives* 22, 2 (2008), 171–192.

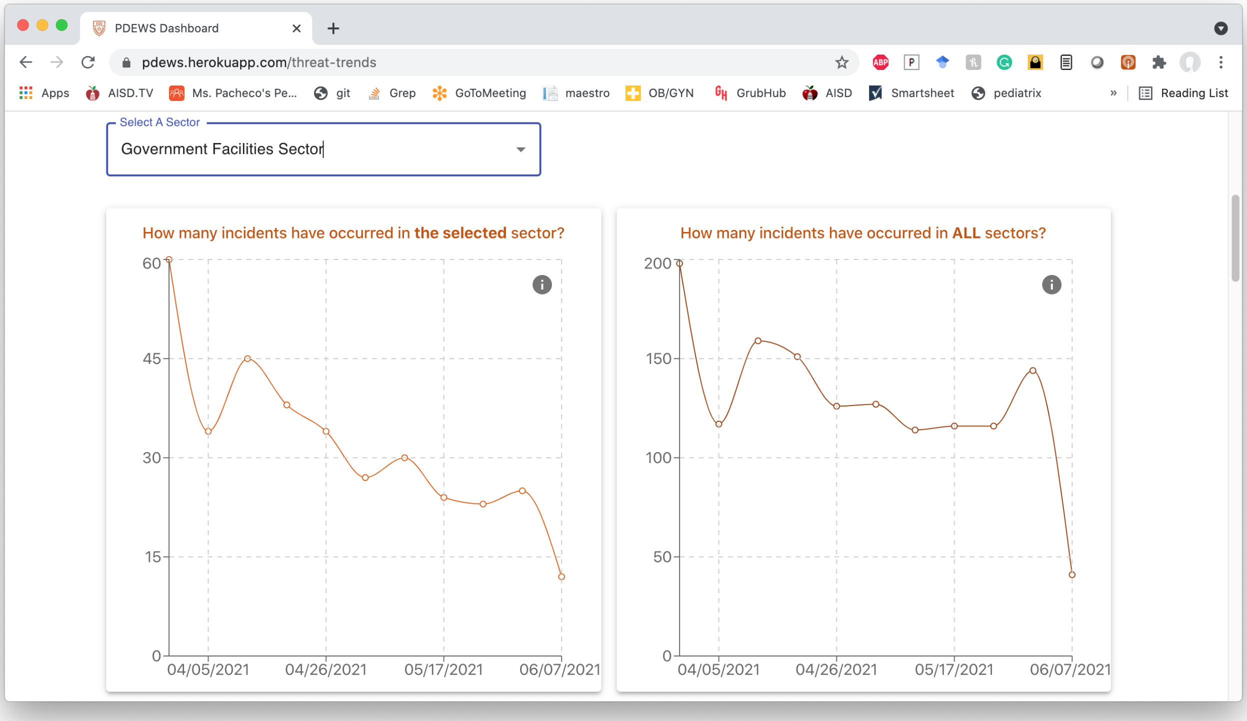
[3] Wade Baker, Mark Goudie, Alexander Hutton, C David Hylender, Jelle Niemantsverdriet, Christopher Novak, David Ostertag, Christopher Porter, Mike Rosen, Bryan Sartin, et al. 2011. 2011 data breach investigations report. *Verizon RISK Team, Available: www. verizonbusiness. com/resources/reports/rp_databreach-investigationsreport-2011_en_xg. pdf* (2011), 1–72.

[4] Ram Basnet, Srinivas Mukkamala, and Andrew H Sung. 2008. Detection of phishing attacks: A machine learning approach. In *Soft computing applications in industry*. Springer, 373–383.

[5] Axton Betz-Hamilton. 2020. A Phenomenological Study on Parental Perpetrators of Child Identity Theft. *Journal of Financial Counseling and Planning* (2020).

[6] Siddhartha Bhattacharyya, Sanjeev Jha, Kurian Tharakunnel, and J Christopher Westland. 2011. Data mining for credit card fraud: A comparative study. *Decision Support Systems* 50, 3 (2011), 602–613.

[7] Rupa Ch, Thippa Reddy Gadekallu, Mustufa Haider Abidi, and Abdulrahman Al-Ahmari. 2020. Computational system to classify cyber crime offenses using machine learning. *Sustainability* 12, 10 (2020), 4087.



Select A Sector: Government Facilities Sector

Figure 3: A screen-shot of the PDEWS front-end: The number of incidents charts for the Commercial Facilities Sector.

[8] Kai Chih Chang, Razieh Nokhbeh Zaeem, and K Suzanne Barber. 2018. Enhancing and evaluating identity privacy and authentication strength by utilizing the identity ecosystem. In *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*. 114–120.

[9] Kai Chih Chang, Razieh Nokhbeh Zaeem, and K Suzanne Barber. 2018. Internet of Things: securing the identity by analyzing ecosystem models of devices and organizations. In *2018 AAAI Spring Symposium Series*.

[10] Kai Chih Chang, Razieh Nokhbeh Zaeem, and K Suzanne Barber. 2020. A Framework for Estimating Privacy Risk Scores of Mobile Apps. In *International Conference on Information Security*. Springer, 217–233.

[11] Kai Chih Chang, Razieh Nokhbeh Zaeem, and K Suzanne Barber. 2020. Is Your Phone You? How Privacy Policies of Mobile Apps Allow the Use of Your Personally Identifiable Information. In *2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*. IEEE, 256–262.

[12] Francisco Charte, Antonio J Rivera, María J del Jesus, and Francisco Herrera. 2015. MLSMOTE: approaching imbalanced multilabel learning through synthetic instance generation. *Knowledge-Based Systems* 89 (2015), 385–397.

[13] Chia-Ju Chen, Razieh Nokhbeh Zaeem, and K Suzanne Barber. 2019. Statistical analysis of identity risk of exposure and cost using the ecosystem of identity attributes. In *2019 European Intelligence and Security Informatics Conference (EISIC)*. IEEE, 32–39.

[14] Hsinchun Chen, Wingyan Chung, Jennifer Jie Xu, Gang Wang, Yi Qin, and Michael Chau. 2004. Crime data mining: a general framework and some examples. *Computer* 37, 4 (2004), 50–56.

[15] Consumer Sentinel Network. 2019. Data Book for January - December 2019. https://www.ftc.gov/system/files/documents/reports/consumer-sentinel-network-data-book-2019/consumer_sentinel_network_data_book_2019.pdf

[16] Heith Copes and Lynne M Vieraitis. 2009. Understanding identity theft: Offenders' accounts of their lives and crimes. *Criminal Justice Review* 34, 3 (2009), 329–349.

[17] NV Gemalto. 2017. Mining for Database Gold: Findings from the 2016 Breach Level Index.

[18] Erika Harrell, Bureau of Justice Statistics, US Dept of Justice, and Office of Justice Programs. 2015. Victims of Identity Theft, 2014.

[19] Chris Jay Hoofnagle. 2007. Identity theft: Making the known unknowns known. *Harv. JL & Tech.* 21 (2007), 97.

[20] Xiaochen Hu, Xudong Zhang, and Nicholas P Lovrich. 2020. Forecasting identity theft victims: Analyzing characteristics and preventive actions through machine learning approaches. *Victims & Offenders* (2020), 1–30.

[21] Identity Theft Resource Center. 2019. Data Breaches. http://www.idtheftcenter.org/id-theft/data-breaches.html

[22] Markus Jakobsson and Steven Myers. 2006. Phishing and countermeasures: understanding the increasing problem of electronic identity theft. *John Wiley & Sons* (2006).

[23] Javelin Strategy and Research. 2014. Identity Fraud Report: Card data breaches and inadequate consumer password habits fuel disturbing fraud trends.

[24] David Liau, Razieh Nokhbeh Zaeem, and K Suzanne Barber. 2019. Evaluation framework for future privacy protection systems: a dynamic identity ecosystem approach. In *2019 17th International Conference on Privacy, Security and Trust (PST)*. IEEE, 1–3.

[25] David Liau, Razieh Nokhbeh Zaeem, and K Suzanne Barber. 2020. A Survival Game Analysis to Personal Identity Protection Strategies. In *2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*. IEEE, 209–217.

[26] Desla Mancilla and Jackie Moczygemba. 2009. Exploring medical identity theft. *Perspectives in health information management/AHIMA, American Health Information Management Association* 6, Fall (2009).

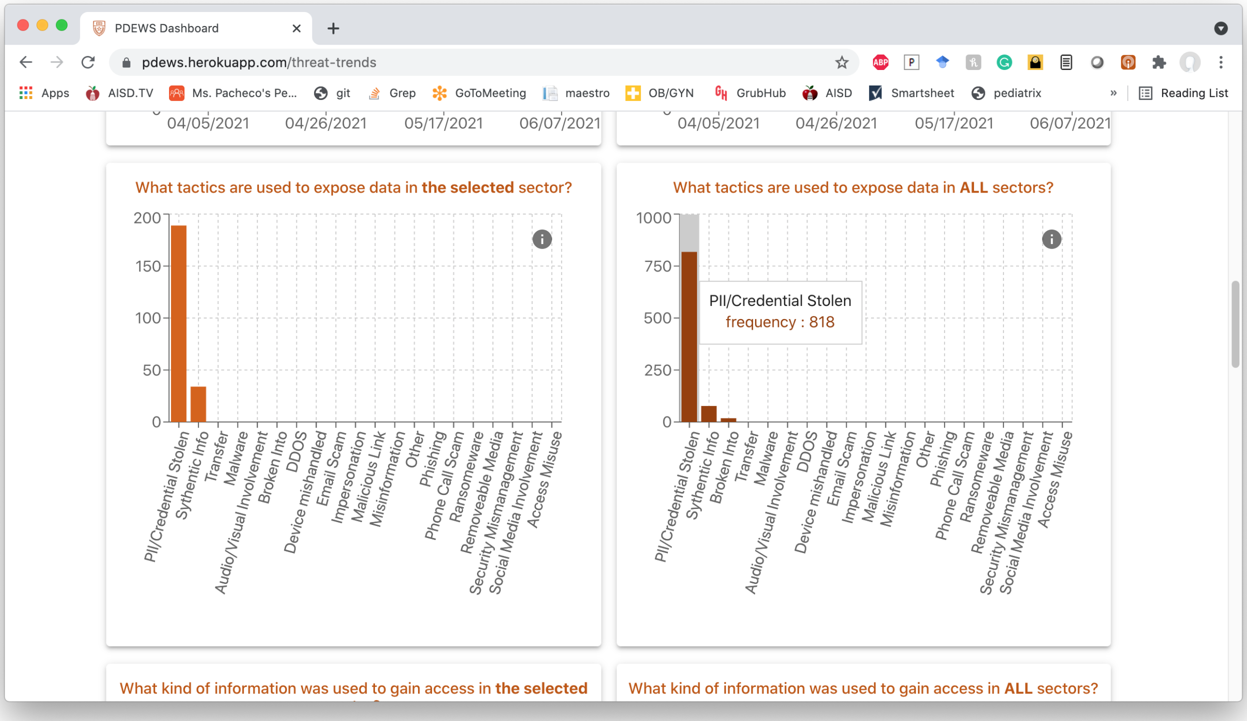
[27] George R Milne. 2003. How well do consumers protect themselves from identity theft? *Journal of Consumer Affairs* 37, 2 (2003), 388–402.

[28] Robert G Morris. 2010. Identity thieves and levels of sophistication: Findings from a national probability sample of American newspaper articles 1995–2005. *Deviant Behavior* 31, 2 (2010), 184–207.

[29] Graeme R Newman and Megan M McNally. 2005. Identity theft literature review. *United States Department of Justice: National Institute of Justice* (2005).

[30] Privacy Rights Clearinghouse. 2019. Privacy Rights Clearinghouse. https://www.privacyrights.org

[31] Darren Quick and Kim-Kwang Raymond Choo. 2014. Data reduction and data mining framework for digital forensic evidence: storage, intelligence, review and archive. *Trends & Issues in Crime and Criminal Justice* 480 (2014), 1–11.



What tactics are used to expose data in **the selected** sector?

What tactics are used to expose data in **ALL** sectors?

What kind of information was used to gain access in **the selected**

What kind of information was used to gain access in **ALL** sectors?

Figure 4: A screen-shot of the PDEWS front-end: The tactics charts for the Commercial Facilities Sector.

[32] Darren Quick and Kim-Kwang Raymond Choo. 2016. Big forensic data reduction: digital forensic images and electronic evidence. *Cluster Computing* (2016), 1–18.

[33] Rima Rana, Razieh Nokhbeh Zaeem, and K Suzanne Barber. 2018. Us-centric vs. international personally identifiable information: a comparison using the UT CID identity ecosystem. In *2018 International Carnahan Conference on Security Technology (ICCST)*. IEEE, 1–5.

[34] Rima Rana, Razieh Nokhbeh Zaeem, and K Suzanne Barber. 2019. An Assessment of Blockchain Identity Solutions: Minimizing Risk and Liability of Authentication. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. 26–33.

[35] Aditya Tyagi, Razieh Nokhbeh Zaeem, and K. Suzanne Barber. 2021. EarlyWarning Identity Threat and Mitigation System. In *1st International Workshop on Cyber Forensics and Advanced Threat Investigations in Emerging Technologies (CFATI3)*. Under Submission.

[36] United States Department of Justice. 2019. National Crime Victimization Survey: Identity Theft Supplement, 2016. https://doi.org/10.3886/ICPSR36829.v1

[37] N. S. Van der Meulen. 2011. Betweeen awareness and ability: Consumers and financial identity theft. *Communications & Strategies* 81 (2011), 23–44.

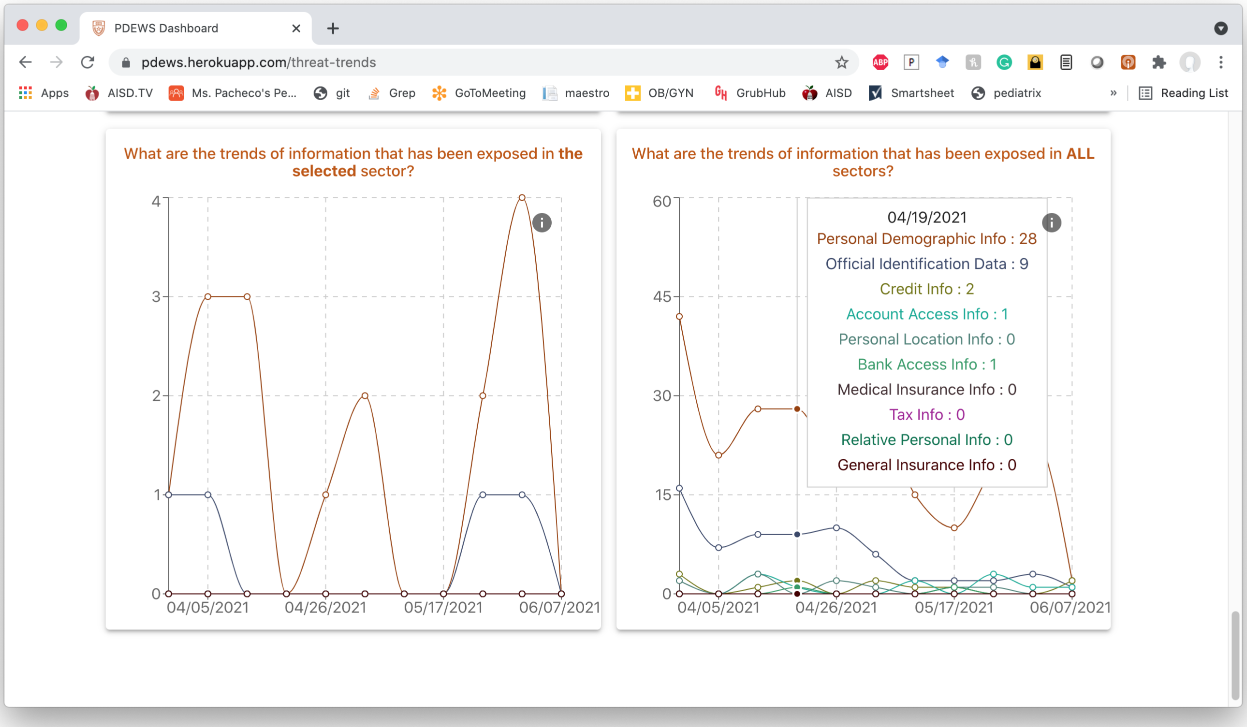
[38] Razieh Nokhbeh Zaeem and K Suzanne Barber. 2020. How Much Identity Management with Blockchain Would Have Saved Us? A Longitudinal Study of Identity Theft. In *International Conference on Business Information Systems*. Springer, 158–168.

[39] Razieh Nokhbeh Zaeem, Suratna Budalakoti, K Suzanne Barber, Muhibur Rasheed, and Chandrajit Bajaj. 2016. Predicting and explaining identity risk, exposure and cost using the ecosystem of identity attributes. In *2016 IEEE International Carnahan Conference on Security Technology (ICCST)*. IEEE, 1–8.

[40] Razieh Nokhbeh Zaeem, Monisha Manoharan, and K Suzanne Barber. 2016. Risk kit: Highlighting vulnerable identity assets for specific age groups. In *2016 European Intelligence and Security Informatics Conference (EISIC)*. IEEE, 32–38.

[41] Razieh Nokhbeh Zaeem, Monisha Manoharan, Yongpeng Yang, and K Suzanne Barber. 2017. Modeling and analysis of identity threat behaviors through text mining of identity theft stories. *Computers & Security* 65 (2017), 50–63.

[42] Jim Zaiss, Razieh Nokhbeh Zaeem, and K Suzanne Barber. 2019. Identity Threat Assessment and Prediction. *Journal of Consumer Affairs* 53, 1 (2019), 58–70.



What are the trends of information that has been exposed in **the selected** sector?

What are the trends of information that has been exposed in **ALL** sectors?

04/19/2021
Personal Demographic Info : 28
Official Identification Data : 9
Credit Info : 2
Account Access Info : 1
Personal Location Info : 0
Bank Access Info : 1
Medical Insurance Info : 0
Tax Info : 0
Relative Personal Info : 0
General Insurance Info : 0

Figure 5: A screen-shot of the PDEWS front-end: The trends of exposed information charts for the Commercial Facilities Sector.

WWW.IDENTITY.UTEXAS.EDU