COMPUTERS & SECURITY 65 (2017) 50–63

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

Modeling and analysis of identity threat behaviors through text mining of identity theft stories

Razieh Nokhbeh Zaeem *, *Monisha Manoharan*, *Yongpeng Yang*, *K. Suzanne Barber*

The Center for Identity, The University of Texas at Austin, USA

ARTICLE INFO

Article history:
Received 21 February 2016
Received in revised form 27 September 2016
Accepted 4 November 2016
Available online 9 November 2016

Keywords:
Identity theft
Text mining
News stories
PII attributes
Named entity recognition

ABSTRACT

Identity theft, fraud, and abuse are problems affecting the entire society. Identity theft is often a "gateway" crime, as criminals use stolen or fraudulent identities to steal money, claim eligibility for services, hack into networks without authorization, and so on. The available data describing identity crimes and their aftermath are often in the form of recorded stories and reports by the news press, fraud examiners, and law enforcement. All of these sources are unstructured. In order to analyze identity theft data, this research proposes an approach which involves the novel collection of online news stories and reports on the topic of identity theft. Our approach pre-processes the raw text and extracts semi-structured information automatically, using text mining techniques. This paper presents statistical analysis of behavioral patterns and resources used by thieves and fraudsters to commit identity theft, including the identity attributes commonly linked to identity crimes, resources thieves employ to conduct identity crimes, and temporal patterns of criminal behavior. Furthermore, the automatically extracted information is validated against manually investigated news stories. Analyses of these results increase empirical understanding of identity threat behaviors, offer early warning signs of identity theft, and thwart future identity theft crimes.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

Identity theft is an ever-present and growing issue in society. According to the National Institute of Justice (Newman and McNally, 2005), identity theft has become perhaps the defining crime of the information age. Identity theft is a crime that affects many people, businesses, and agencies. In 2013, the number of identity theft victims exceeded 13 million, with economic losses totaling $18 billion (Javelin Strategy and Research, 2014). In the same year, the Federal Trade Commission (FTC) received more than 290,000 complaints related to identity theft, the highest number of identity theft complaints ever received by the FTC (Federal Trade Commission, 2016a). In addition, identity theft is an issue of global concern. Identity-related crimes accounted for more than 60% of confirmed fraud reports in the United Kingdom (CIFAS, 2014), and an Australian Fraud report noted that 19.8% of people surveyed experienced two or more incidents of identity fraud within the previous five years (Australia Bureau of Statistics, 2011). Identity theft is a costly problem with constantly evolving patterns of criminal tactics and behaviors (Van der Meulen, 2011).

* *Corresponding author.*
E-mail addresses: razieh@identity.utexas.edu (R. Nokhbeh Zaeem), monisha@utexas.edu (M. Manoharan), yangyp@utexas.edu (Y. Yang), sbarber@identity.utexas.edu (K. S. Barber).
http://dx.doi.org/10.1016/j.cose.2016.11.002
0167-4048/© 2016 Elsevier Ltd. All rights reserved.

As more businesses and people become victims of identity crimes, it is increasingly important to better understand the crimes of identity theft, fraud, and abuse in an effort to reduce or even halt this crime. While statistics have been gathered regarding the number of exposed records or the financial loss to individuals (Identity Theft Resource Center, 2014; Privacy Rights Clearinghouse, 2016), few efforts have researched how identity theft actually occurs. There are best practices and prevention tips from security companies and government agencies available (Federal Trade Commission, 2016b; The United States Department of Justice, 2016; CSID, 2016; Lifelock, 2016). However, there is a lack of aggregated data detailing the process of stealing someone's identity. Most information available focuses on reactive measures, which are helpful once your identity is stolen, but brings us no closer to ending identity-related crimes or, at least, making identity theft more difficult for the criminal. The consumers are several steps behind the identity thieves.

The aim of this project is to develop a repository of relevant knowledge to better understand the processes used by identity thieves and fraudsters. Our aim is to understand the criminals' process, the vulnerabilities that allow the crime to take place, the resources that facilitate it, and what can be done to prevent it. We hope this knowledge will bring about a shift in the definition and use of credentials to decrease identity theft and fraud vulnerabilities. To achieve this goal, the Center for Identity at The University of Texas at Austin built the Identity Threat Assessment and Prediction (ITAP) tool (AWARE Software Inc., 2014) to describe the business-like model of criminal methods and develop techniques to better analyze crimes that steal and use identity information. ITAP gives us a better understanding of fraudsters' behaviors and patterns based on past identity theft and fraud. ITAP will deliver actionable knowledge grounded in the study of thefts that have actually happened in the past. ITAP seeks to answer relevant research questions, such as: How are these perpetrators gathering information? What resources are being used to overcome security hurdles? What process steps are being taken to steal someone's identity?

In order to assess and predict identity threats, analytical tools like ITAP need sufficient amounts of data on which to perform analysis and generate reliable results and conclusions. The main issue, however, is that there is no publicly available repository describing ongoing identity theft and fraud. Second, it is hard to keep this database up-to-date as the magnitude of identity theft and fraud increases each day. Thus, a lack of identity theft data in a structured form is a problem that limits the ability to model identity theft. Although some government organizations have internal databases cataloging identity theft and fraud, these are not readily available to the public or to research institutions. Another possible source is the Internet, which publishes a multitude of identity theft news stories. However, these articles are in a raw text format and cannot be analyzed directly to find patterns. In this work, we attempt to solve this problem by proposing an automated solution that uses text mining, an application of natural language processing, to extract the useful information from the identity theft stories and articles. This information can then be used to extract patterns in identity thief behaviors and offer strategies to prevent future identity theft crimes. This work is an initial effort toward identity theft modeling. Our ongoing research efforts seek to evolve the model to a point where it can become a publicly available tool that helps inform businesses, consumers, agencies, and researchers of the processes of identity theft.

This paper makes the following contributions:

1. Novel application of natural language processing to collect information about habits and methods of identity thieves;
2. First time use of news stories to study risks, losses, and trends in identity theft, applied using over 3500 news articles;
3. Comparison with more than 250 manually investigated identity theft stories.

2. Sources of identity theft information

Previous work has introduced several avenues of collecting identity theft data. In this section, we briefly cover these sources and introduce our novel use of news stories.

2.1. *Agency data*

The first commonly used source of identity theft data is gathered from agencies (Allison et al., 2005; Federal Trade Commission, 2014). For example, the FTC's Consumer Sentinel Network (Federal Trade Commission, 2014), established in 1997, is a database which collects identity theft complaints from FTC's telephone- and web-based complaint systems in addition to more than one hundred federal and state organizations. While comprehensive (Consumer Sentinel Network has more than 10 million complaints as of 2014), this database is available only to law enforcement. Even though some have raised concerns (Newman and McNally, 2005) toward the representativeness of Consumer Sentinel Network and other agency data, researchers still use agency data widely, while making an effort to manage the amount of data in such databases (Quick and Choo, 2014, 2016).

2.2. *Surveys*

Synovate on behalf of the FTC, Javelin Strategy and Research (Javelin Strategy and Research, 2014), 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 (Newman and McNally, 2005) 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.

2.3. *Anecdotal information*

One other source of data about identity theft is through victim case studies. While this source highlights the worst possible scenarios, it is the least reliable and most biased.

2.4. *Our contribution: news stories*

News stories have been used in general for text mining (Dolan et al., 2004; Erickson and Howard, 2007) and in particular for

identity theft (Morris, 2010; Morris and Longmire, 2008). However, the application of news articles to extract statistics *about* identity theft is novel (Yang et al., 2014).

News stories have several important characteristics that make them an appropriate source of data:

1. Volume: There is a tremendous amount of news stories about identity theft.
2. Availability: One of the problems in the identity research area is that it is difficult to obtain source data from government agencies or corporations. However, the information published in a news story is publicly available. Therefore, researchers need not worry about protecting the Personally Identifiable Information attributes associated with reported victims.
3. Reliability: Identity theft news stories are mostly reliable. Although some information published on the Internet is not accurate or even false, most news stories are reliable and trustworthy since the news media is responsible for providing accurate information to the public and are held accountable by editors and their viewers.

News stories complement 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 (Golden, 2013). For example, an identity theft resulting in a small monetary loss may not be considered newsworthy, and hence is not reported and not included when calculating the average monetary loss per incident. Thus, the average calculated losses may be higher than the ground truth value.

3. Background of text mining

The problem of information extraction from raw text format has been approached by numerous research efforts. However, application of this technique in the identity research area is relatively novel. This paper is based on the Identity Threat Assessment and Prediction (ITAP) project at the Center for Identity at The University of Texas at Austin and related works in the natural language processing research and text mining studies.

Text mining refers to the process of gleaning meaningful information from natural language text. The goal is to analyze the text and extract useful information for a specific use (Witten et al., 2004). It is essentially an application of natural language processing to transform the natural text into directly usable data (Manning and Schütze, 1999). Unlike the well-formed data stored in a database, natural language text is unstructured and difficult to understand by computers. Thus, text mining usually requires transforming the natural language text into a structured format, detecting lexical and syntactic usage patterns, and evaluating and analyzing the generated data. Typical text mining research includes text categorization, entity extraction, sentiment analysis, entity relation modeling, and so on (Sanger and Feldman, 2006). Text mining techniques help process large amounts of unstructured data, such as biomedical applications (e.g. association of gene clusters), social network applications (e.g. social network hashtag trends), marketing applications (e.g. customer relationship management), and sentiment analysis (e.g. customer sentiments on movies) (Aase, 2011). This section briefly introduces text mining techniques, with a focus on the methods used in this project.

3.1. *Text preprocessing*

Before actually analyzing the natural language text, preprocessing is usually done to eliminate the language-dependent factors so that the language structure becomes more clear (Wang and Wang, 2005). Tokenization is one of the most common techniques used for text preprocessing. It refers to the process of splitting a text stream, such as a sentence, into tokens, such as phrases, words, symbols or other kind of elements. In the text mining field, a token usually means a sequence of characters that are classified together as a group to represent a meaningful semantic unit for processing (Al and Abu, 2013). The commonly used approach simply splits the text or sentences based on white spaces, punctuation, or other special symbols between words. After tokenization, stop-word removal and lemmatization may be applied for further preprocessing (Sanger and Feldman, 2006). Stop-words refer to high frequency words in a language that do not carry any significant meaning, such as the articles “the”, “a”, “an”, etc. For a specific application domain, one can also create stop-word lists by applying statistical measures to remove the less informative words. Removing these stop-words helps reduce noise and select meaningful textual features. Lemmatization is the process of reducing inflected words into a base form so that the number of phrases or words with similar meaning is reduced. For example, English words like “look” can be inflected with a morphological suffix to produce similar words such as “looks”, “looking”, and “looked”. These words all share the base “look”. It is usually beneficial to map all inflected forms into the base.

3.2. *Noun phrases*

Noun phrases (NP) are units whose first or principal word is a noun, pronoun or other noun-like words, which can be modified by words such as adjectives (Serrano and Araujo, 2005). Noun phrases are the main carriers of the content of a text document and can be used to extract more informative features and meaningful information than a single word. For example, “social security number” is a noun phrase. Proper nouns, which are a subset of noun phrases, are the nouns that represent unique entities (Lester and Beason, 2012), such as “San Francisco”, “LeBron James”, or “Philadelphia 76ers”. These words are distinguished from the common nouns that refer to a class of entities or non-unique instances of a certain class, such as “person” or “these persons”.

3.3. *Named entity recognition*

Named entities are phrases that contain the names of persons, organizations, locations, expressions of times, monetary values, and so on. For example: “James watched an NBA game last

night". This sentence contains three named entities: "James" is a person, "NBA" is an organization, and "last night" is time. Named entity recognition (NER) is an important task in information extraction, which locates and classifies the words or phrases into predefined categories (Tjong et al., 2003). NER systems have been implemented by using a variety of models, such as Hidden Markov models, Maximum Entropy Markov models, and Conditional Random Fields (CRF) (Lafferty et al., 2001). Stanford's Natural Language Processing (NLP) Group has developed a new approach that incorporates non-local structure to augment an existing CRF-based information extraction system with long-distance dependency models, which reduces the error up to nine percent over state-of-the-art systems (Finkel et al., 2005).

3.4. *Part-Of-Speech Tagging*

Part-Of-Speech (POS) Tagging is a process that assigns a word or a phrase in a text to a corresponding POS tag, such as noun, verb, adjective, and so on. This process is based on both the definition of the word as well as its context, which means the same word can have different tags with different adjacent or related words. For example, the word "record" could be a noun or a verb depending on the particular context. The two most common approaches for POS tagging are rule-based and learning-based. The rule-based approach is based on human crafted rules using lexical and other linguistic knowledge. The learning-based approach trains the model based on human annotated texts such as the Penn Treebank (Marcus et al., 1993). The learning-based approach has proven to be more effective considering the devoted human effort and expertise. The Stanford POS tagger used in this project is built by the Stanford NLP group, which combines multiple features with a Cyclic Dependency Network that has 97.24% accuracy on the Penn Treebank WSJ, reducing the error by 4.4% compared to the best previous single automatically learned tagging result (Toutanova et al., 2003).

4. Proposed algorithm

This section explains the algorithms and design used to mine the news articles gathered from the Internet. The idea is to design a pipe-lined system that takes identity theft news stories from the Internet as input and generates the analytics that help us better understand the identity theft process as output.

4.1. *News articles collection*

The first step in the data collection process is to get the news article links from the Internet. We gathered the articles by (1) using search engines and searching for a set of key words highly related to identity theft and (2) from annual identity theft reports.

4.1.1. *Searching for identity theft stories*

We manually constructed a list of key words for the search. Table 1 lists the news Rich Site Summary (RSS) URLs obtained by searching the key words on Google and New York Times News. Each Google news RSS contains 100 original links to identity theft news stories, which we used to collect the stories daily. The New York Times News RSS also provides several stories each day. Worth mentioning here is that some of the news stories collected by this method may not be identity theft victim reports as expected. For example, an article explaining how to protect yourself from identity theft could be gathered using this search.

Table 1 – News RSS URLs based on identity theft related keywords.

NEWS sources	URL
Google News	https://news.google.com/news/feeds?q=identity+theft&num=100&output=rss
Google News	https://news.google.com/news/feeds?q=identity+thieves&num=100&output=rss
Google News	https://news.google.com/news/feeds?q=identity+fraud&num=100&output=rss
New York Times	http://topics.nytimes.com/top/reference/timestopics/subjects/i/identity_fraud/?rss=1

4.1.2. *Identity Theft Resource Center*

Additionally, links are extracted from publicly available annual identity theft reports on The Identity Theft Resource Center (2014). Their breach report consists of data breaches gathered from a variety of media sources and lists from state governmental agencies.

Then, links pointing to invalid URL addresses are discarded and duplicate links are eliminated. While about 300 stories are posted each day on average, around 40 valid stories are obtained after eliminating the duplicate and invalid links. In order to keep track of the source, the original links to the articles are stored along with the stories. The data in this study consist of approximately 3500 valid articles.

4.2. *Pipe-lined system model*

The pipe-lines system is composed of the following steps, as shown in Fig. 1:

1. With the news stories from the Internet as input, the first step is to get their main textual content. In order to extract news articles from the published format (usually HTML), we used the boilerpipe library (Boilerplate, 2016) and removed format labels, navigation lists, advertisements, etc. Moreover, sometimes the links point to PDF files that contain the article instead of HTML files. For such articles, we use a PDF extractor, developed based on the PDFBOX library (Apache, 2016), to extract the stories and store the content in a text file.
2. Next, the story text is preprocessed; unused language-dependent factors are eliminated. We used tokenization (PTBTokenizerAnnotator from the Stanford CoreNLP library (Finkel et al., 2005)) and lemmatization. The PTBTokenizerAnnotator is a PTB (Penn Treebank) style tokenizer.

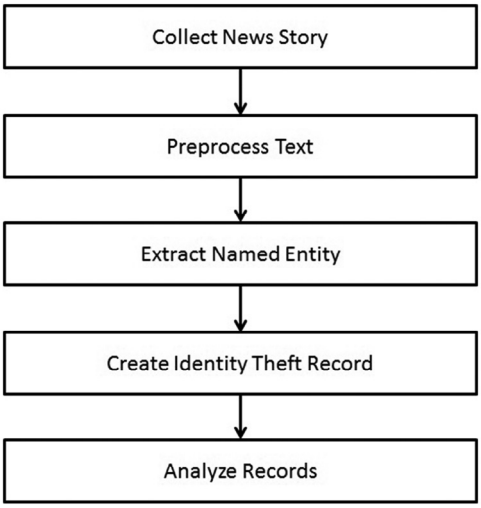

Fig. 1 – Pipe-lined system model.

3. Then, named entities are extracted using the named entity recognizer. From the Stanford CoreNLP library, we used Named Entity Recognizer (NER) and Part-Of-Speech (POS) tagger functions.
4. Next, each story is represented as an identity theft record – a predefined representation of an identity theft crime. Each named entity (word or phrase) in the news story is assigned by the CoreNLP library to a corresponding category, such as people, organizations, money, and time. This assignment is saved in a DAT file that maps various entries of a record such as victim, organization, location, date, cost, resource, actions, etc. to their values in the current story. This DAT file serves as the identity theft record.
5. This record is then used to conduct analysis about different aspects of the identity theft.

The following sections elaborate on details of step 4, which builds the identity theft record.

4.2.1. *Defining the PII attributes*

In an identity theft record, an entry belongs to Personally Identifiable Information (PII) attributes stolen or used as resources by the thieves. A pre-existing list of PII attributes (Table 2) is defined manually by selecting the identity attributes commonly used by identity thieves. The words in the pre-existing attributes list are compared against the words in the news articles. Matched words are stored into the corresponding identity theft story record. For example, the attribute social security number is a predefined PII attribute. ITAP looks up social security number in the articles when processing the new stories and counts its frequency of occurrence. Such information is stored and used later to conduct further analysis.

4.2.2. *Impacted target*

We define an impacted target as a person or an organization that has experienced an identity theft or fraud. The identity theft impacted target could be an individual person, a corporation, or a state or federal government agency. The NER recognizes names, possibly more than one name per story, and categorizes them to the corresponding types – individuals, corporations, state government agencies, and federal government agencies. If a corporation with a certain number of individuals is breached, then it is counted as one corporation because there are stories that do not report the number of individuals involved.

Table 2 – Complete list of PII used.

PII attribute	
Utility account	Bank account
Vendor name	Health insurance card
Military ID	Social security card
Medicare card	Credit score
Timezone	Transaction flags
Vendor number	Maiden name
Balance sheet	Account number
Loyalty card	Military rank
Phone number	Medicaid card
Employee ID	Medical history
Credit rating	Zip code
Business type	Operating system
Stock price	Blood type
Social security number	Facebook account
Birth certificate	Spouse name
Credit/debit card	Access card
Email	Application name
Personal information	Administrator password
Insurance ID	Stock exchange ticker symbol
Student ID	

4.2.3. *Risk calculation*

Risk is calculated for each PII attribute based on its frequency of occurrence in the news article. The predefined PII attribute list is used here. For each predefined PII attribute, the occurrence is counted as the news story is processed. We hypothesize that the frequency of occurrence for a PII attribute implies the probability the attribute is exposed in the identity theft process. Thus, an attribute with higher frequency of occurrence in the news story will have a higher risk of exposure.

4.2.4. *PII category*

Identity theft happens across a variety of areas. The PII category here refers to the areas or industries where the identity theft occurred. We manually assigned each of the PII attributes to a category. Examples of these categories include:

1. What You Know: the resources/attributes that an individual has knowledge about, such as one's home address and mother's maiden name.
2. What You Are: the attributes that are part of the physical being of a person such as DNA and fingerprint.
3. Federal Government: the attributes that are issued by the federal government, such as social security number, passport, and visa.
4. State Government: the attributes assigned by state governments such as state ID card.
5. Bank: the attributes assigned by banks, such as bank account number.
6. Consumers Services: the attributes assigned by consumer services companies, such as rewards program number.

7. Medical: the attributes assigned and used in the medical industry such as health insurance information.

4.2.5. *Time selection*

The time of occurrence of an identity theft is important to produce correct analysis, such as calculating the loss and risk changes in reference to time. Due to the non-deterministic nature of the duration of an identity theft and the delay in publishing of the news story, it is hard to get an accurate time measurement. Two approaches could be used to analyze approximate time from the news articles. The first approach is to simply record the time that the articles are published. The problem with this approach is that the identity thefts may have happened long before the news articles are actually published. This could result in extracting a time that does not accurately reflect the date of the identity theft. The second method is to record the time obtained from the news story by using the named entity recognizer. In this way, the time of each identity theft is decided by the content of each article and is highly related to the theft itself. This approach works in most cases, but some special conditions must be considered. For example, there might be multiple dates mentioned in the article since the total process may last for several months. To deal with such conditions, the ITAP collects all valid dates as the time the theft happened and weighs them equally when used later for further analysis.

4.2.6. *Finding the location*

Location statistics can determine the states which are at most risk for identity theft and the accumulative losses in those states. The location details captured by the named entity recognizer are mostly accurate and are directly stored into the Location entry in the DAT file. There are some similar problems as encountered in the time selection, such as multiple instances of location and missing location information. These problems are handled in a similar way as done for the time selection. Multiple locations are weighted equally and the location entry for missing locations is marked as not available.

4.2.7. *Loss calculation*

Loss is calculated based on the occurrences of the named entity "money" in the news article. The format of the loss obtained by the named entity recognizer could appear in several different ways. A format transformation is necessary to get a unified result. For example, a loss of 1 million dollars may be represented as "$1,000,000" or "$1000000" or just "1 million dollars" in the news stories. To simplify subsequent calculations, the loss is transformed to the form of only digits. Then, all the monetary losses in a single news story are summed up. This summation is the loss for this particular story. In order to analyze the loss experienced by the exposure of a particular PII attribute, some form of weighting is needed. Here, only the attributes matched in the predefined list of PII are considered. An attribute with higher frequency of occurrence in the news story will have a higher weight, hypothesizing that the attribute frequency in an identity theft story indicates its importance in the theft process analyzed. The formula below is used to calculate the weighted loss for each attribute.

$$Loss_{a_i} = \frac{Loss_{sum} * f_{a_i}}{\sum_{i=0}^{n} f_{a_i}} \quad (1)$$

where a_i is a PII attribute; $Loss_{a_i}$ is the loss caused for a particular PII attribute; and f_{a_i} is the frequency of occurrence for a particular PII attribute.

4.2.8. *Timeline*

In order to observe the trend of identity theft loss over time, we introduce the correlation between the loss and the date. Instead of calculating the loss directly, the date information of the identity theft occurrence is added to the calculation. In other words, for each attribute, the loss is assigned equally to the dates that occur in the news story. The new formula for the loss calculation is as follows:

$$Loss_{a_i,d_i} = \frac{Loss_{sum} * f_{a_i}}{\sum_{i=0}^{n} f_{a_i} * N_d} \quad (2)$$

where a_i is a PII attribute; d_i is the date related to the news story; $Loss_{a_i,d_i}$ is the loss caused by a particular PII attribute and assigned to the date d_i; f_{a_i} is the frequency of occurrence for the PII attribute; and N_d is the total number of dates in the story.

5. Results and analysis

This section describes the results from running the proposed algorithm and analyzing the records for more than 3500 identity theft related news stories collected from various news feeds from February through April 2014. Also, a more in-depth analysis on several different aspects of the results is illustrated. Then, the results and analysis are discussed with regard to understanding the identity theft process and prediction of identity threats in the future.

Although the stories provide a fairly large amount of information, most of the stories are lacking some entries in their records. Thus, those identity theft records are sparse which affects the accuracy of the analysis.

5.1. *Impacted target*

Fig. 2 presents the number of impacted targets in the identity theft stories investigated. Since more than one impacted target per article is possible, the total number of impacted targets is more than the total number of stories. From Fig. 2, we can see that individuals are the most frequent targets of identity thieves. Corporations are also targeted since a successful theft on a corporation could bring the thieves significant financial interest. Government agencies are not targeted as often as individuals and corporations, likely because it is more difficult to steal money from these government agencies, and the level of risk for the thieves is much higher.

5.2. *PII attribute risk analysis*

Different PII attributes have different risks of exposure. The risk calculation is based on the assumption that the risk of an attribute or resource is positively correlated with its frequency

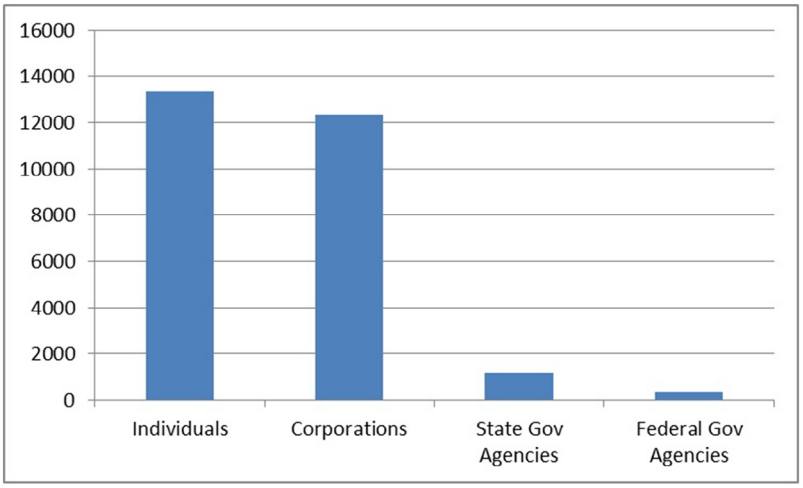

Fig. 2 – Impacted target.

of use. Detecting commonly used resources could help us understand how fraudsters are gaining access to sensitive personal information. Identifying the most vulnerable PII attributes aids in limiting access to them and raising public awareness. Often, these attributes/resources are the pinnacle to the completion of the identity theft process, and ultimately, the theft. Therefore, analysis in this area advances identity theft detection and prevention.

The statistics are shown in Fig. 3. The number in the graph indicates the frequency of occurrence of the attribute/resource in the news stories normalized against the total number of PII's in all the stories. Social Security Number (SSN) is the most used attribute. The SSN attribute is important for almost every aspect of the lives of U.S. citizens. It can be used to apply for driver's licenses, credit/debit cards and various kinds of applications to purchase products or gain access to services. As a form of validation, every other study of identity theft also declares SSN as the most sought after PII attribute by identity thieves. In some scenarios, the thieves even use a victim's SSN and combine it with a name and birth certificate to forge a new identity. Thus, SSN has the highest risk among all the attributes. Credit/debit cards are also under high risk because of their widespread use and direct connection to money.

5.3. *PII category*

Fig. 4 shows the category distribution normalized against the total number of PII's in all the stories. As indicated by Fig. 4, the top four categories – What You Know, Federal Government, Bank and Consumer Services – consist of more than 90% of the total theft investigated. Thus, protecting those areas is important to prevent the loss of individual financial interests.

5.4. *Location analysis*

The location of the identity theft and fraud is also important to explore. Fig. 5 shows the identity theft location-wise distribution in the stories investigated. The figure shows the relative rate of identity theft per capita, with red being the highest rate and yellow being the lowest rate.

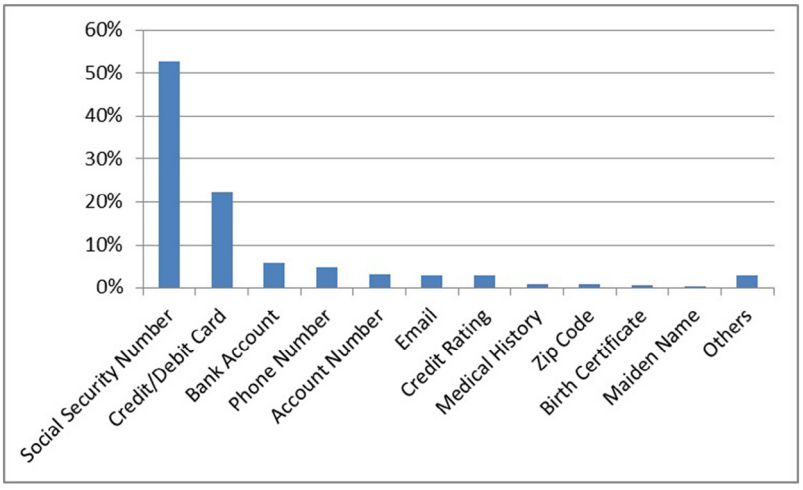

Fig. 3 – PII attribute risk chart.

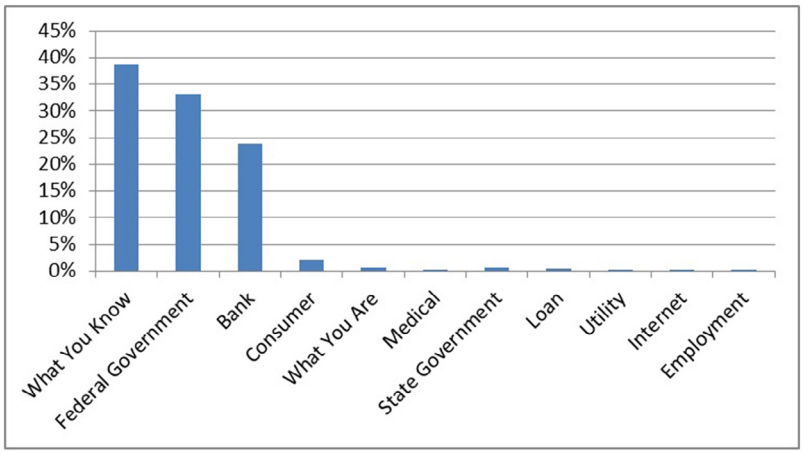

Fig. 4 – PII category distributions.

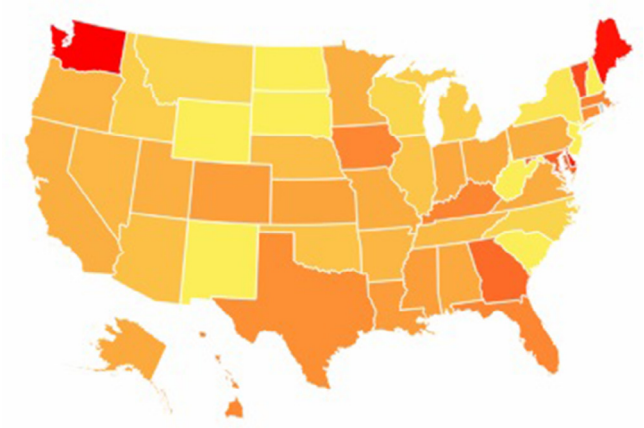

Fig. 5 – Identity theft map.

5.5. *Loss analysis*

The financial impact of identity thefts is of significant concern. Although identity thieves may use the victim's identity to commit non-financial crime, such as launching a terrorist attack, most identity thieves are financially motivated. Which PII attributes, if compromised, can cost the victim the most financial loss? How much should one invest on protecting these attributes? Analyzing such information helps in improving a person's awareness to protect higher valued attributes.

Fig. 6 presents the financial loss caused by theft and fraudulent use of each PII attribute in the identity theft stories investigated. The figure indicates that the results mostly match the attribute risks. Social security number caused the most loss among all attributes, compatible with its high risk. As discussed earlier, this is probably because SSN is important for

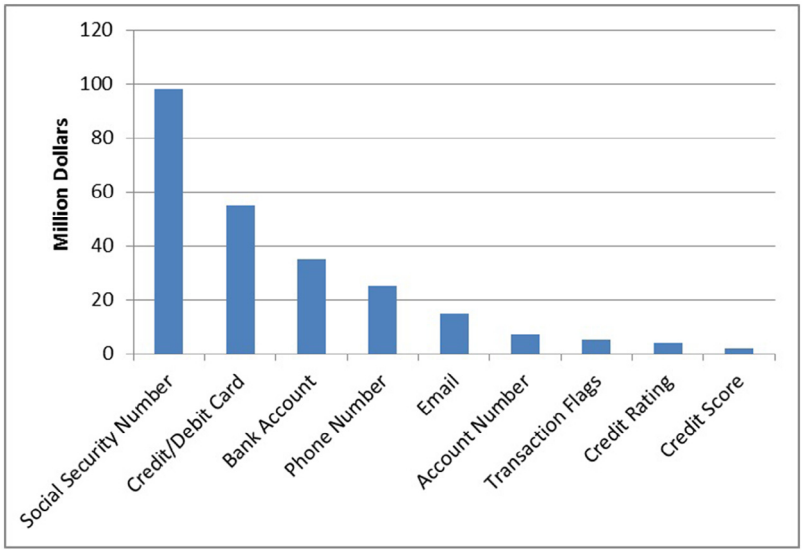

Fig. 6 – Financial impact per attribute.

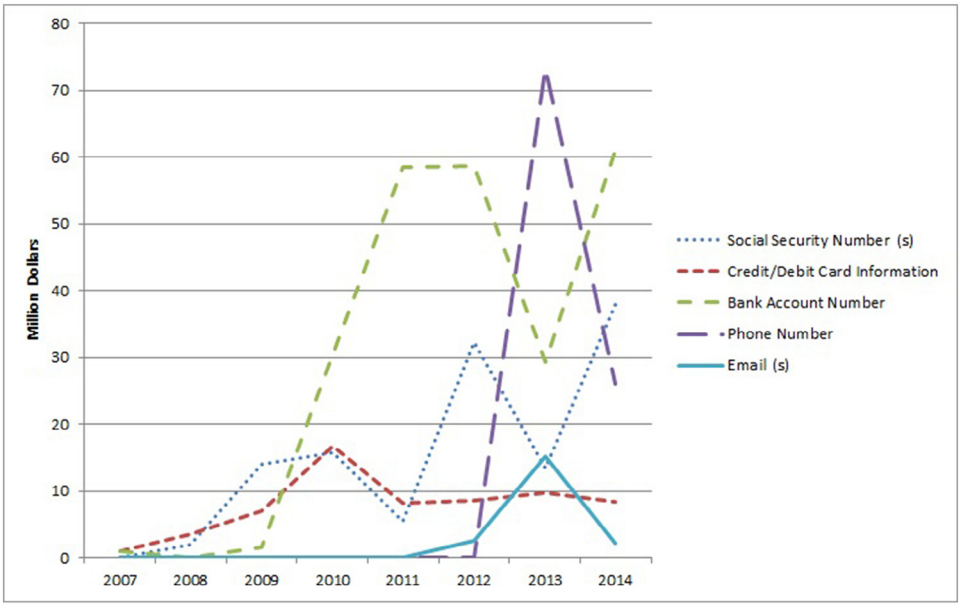

Fig. 7 – Loss per PII attribute per year (Top 5).

many aspects of our lives. The credit/debit card and the bank account are still the second and third most costly attributes. Moreover, the phone number and email here rank fourth and fifth among the most costly attributes, which matches their high risks. At first glance, it seems that the exposure of phone number and email address should not result in so much loss since these PII attributes do not have an explicit financial value and are commonly shared as means of contact. However, considering they are used widely for authentication for financial accounts, the identity thieves could use these attributes to pass the authentication request and reset the password or security questions, which can indirectly lead to access to those accounts and subsequent theft of money.

5.6. *Timeline analysis*

The previous analyses are based solely on the frequency or total loss related to a particular attribute. Time of theft is also an interesting factor. Time trend analysis helps us better understand the trend of change in attribute value. Fig. 7 presents the yearly loss related to the five attributes that caused the most losses in the identity theft stories investigated. The figure shows very few losses before the year 2009. This is because the news stories are mainly collected from the recent years' news.

The figure shows general increase in loss over recent years, bearing in mind that the data from 2014 are only for the first three months. One caveat is that the data set used is biased toward recent thefts, since recent news stories where studied. Having said that, there is other evidence that support the general increase in the amount of identity theft loss by showing that fewer breaches are resulting in more records and financial losses, indicating better targeting by criminals (Teamshatter, 2016). Finally, there is an interesting local drop in loss during 2013 for Bank Account and SSN.

The data used for this research are obtained from gathering the news stories for 45 days (from February 2014 through April 2014) on a daily basis, as well as from the breach report generated by the Identity Theft Resource Center (2014). More data will increase confidence in the results so that the time trend could be used to predict risk and cost trends related to a particular attribute, helping us protect these attributes and get a step ahead of the thieves.

6. Comparison with manual investigation

In this section, we seek to validate the correctness and completeness of our text mining methods by comparing them to manually studied theft stories. As a part of the ITAP project, a team has been looking into close to 5000 identity theft news stories, manually gathering all useful information about each story. We selected all the stories that this team studied and were published from February through April 2014. There were 262 such stories and the team had identified the following properties in each story: PII attributes, performers, victims, location, date of event, date of article, loss, etc., but not impacted target or PII category. Therefore, we compare the findings of the team with our text mining results when data from both manual investigation and text mining methods are available. It is important to note that these are not necessarily the same stories as the input of the text mining method. We only seek to validate the general trend, and this section is not meant to cross validate the results.

Fig. 8 is the counterpart of Fig. 3. Both of these figures show the risk of exposure of most risky PII attributes, calculated based on the frequency of occurrence of PII in the stories, and normalized based on their total. Since the team that manually studied stories enriched their list of PII attributes as they examined new stories, they had a longer list of PII. Once the occurrence numbers normalized, this longer list caused a bigger denominator and hence smaller percentage of risk. However, we can still compare the relative distribution of risks in the

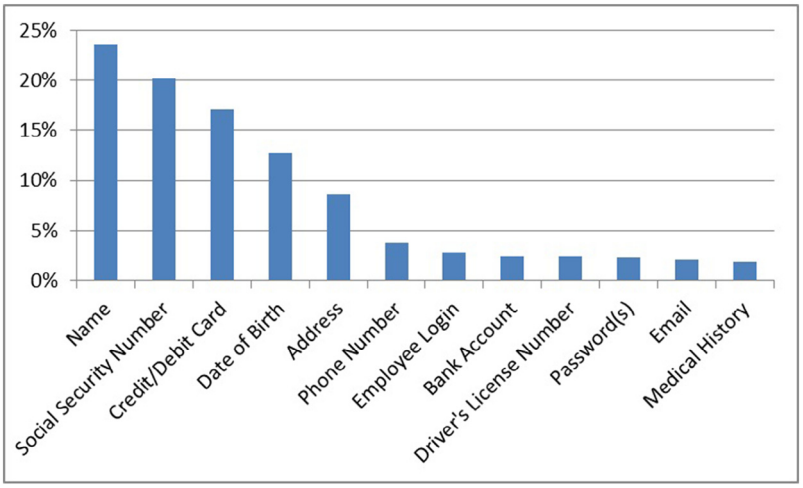

Fig. 8 – PII attribute risk chart (manual investigation).

two figures. According to Fig. 8, the most at risk piece of PII is one's name, which we never included in the PII list of the text mining method as we did not view it as a piece of information that could be stolen. Both figures rank SSN and Credit/Debit Card as the next targeted PII.

Fig. 9, similarly to Fig. 5, shows the location of the examined theft stories. Both figures show the relative frequency of identity theft per capita. We purposefully do not report the absolute per capita numbers, as they are solely based on the studied stories in each method and not the ground truth of identity theft per capita. Washington, Florida, Massachusetts, and Delaware are shared between the top ten states of both manual investigation and text mining methods. Considering the total number of thefts in each state (not per capita), Texas, California, Florida, Washington, and Georgia (in exact order) make the top five states in the text mining approach, and Florida, California, Texas, New York, and Ohio (in exact order) are the top five states in the manual investigation.

Fig. 10 shows the potential loss per attribute. Comparing this figure to Fig. 6 shows that identity theft financial loss is in the order of hundreds of million dollars. However, it appears that in many stories considered by the team the exact PII attribute was not mentioned and was hence identified as Financial Information in general.

Finally, Fig. 11 shows the loss over time of the same PII attributes as of Fig. 7 (not all of them the same as the top five attributes identified in the manual study). Both figures show general increase in the amount of identity theft loss. The same local drop in losses through bank accounts that text mining saw in 2013 is present in the manual examination too.

The results of the manual investigation overall align with the the text mining findings, even though specifics differ. We did not seek to cross validate the text mining method, instead we used the readily available study of the team to validate the overall trends and values.

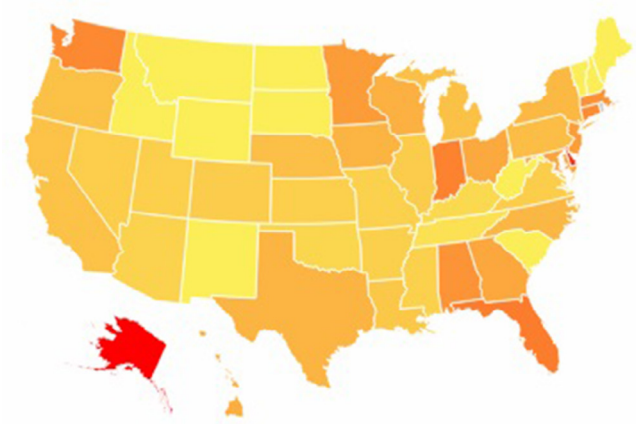

Fig. 9 – Identity theft map (manual investigation).

7. Related work

There exists a large body of research that studies and reports statistics of identity theft. The research is mainly conducted by Federal and State agencies, private organizations, or academic institutions.

From Federal and State agencies, studies by U.S. department of Justice (Harrell, 2015) regularly report the distribution of identity theft victims. Private organizations such as Javelin (Pascual et al., 2016) have published comprehensive studies and case studies on the subject of identity theft and fraud.

In the academia, Allison et al. (2005) found the typical victim of identity theft and fraud to be male and white. Copes et al. (2010) used the National Public Survey on White Collar Crime to separately investigate existing credit card fraud, new credit card fraud, and existing bank account fraud. They reported statistics of victims and compared demographic patterns of identity theft victims to their representation in the general U.S. population. Reyns (2013) performed an in depth empirical examination of identity theft in the United Kingdom from a victimization perspective. Using Routine Activity Theory, he found male older individuals, and those with higher incomes more likely to experience identity theft. Others have performed similar studies based on Routine Activity Theory in various jurisdictions (e.g., Pratt et al., 2010 and Choo, 2011). The

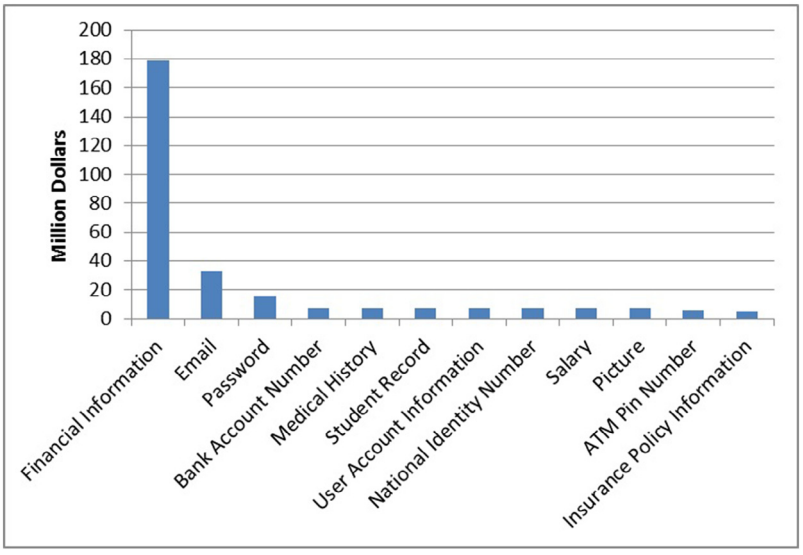

Fig. 10 – Financial impact per attribute (manual investigation).

extraction of input data for these studies, however, is largely manual and does not typically use data mining techniques.

Researchers have applied text mining techniques for a variety of purposes. For example, a recent work (Peng et al., 2016) uses text mining techniques, particularly n-Gram analysis, for behavioral identification of users on social media. Another related research is the mining of unstructured text extracted from the web (e.g., social media) to obtain information (e.g., emergency events and their temporal patterns (Xu et al., 2016a)).

On the subject of identity theft, researchers have leveraged text and data mining in order to prevent and combat identity theft in various ways:

- Facilitate the tasks performed by law enforcement, e.g. Chen et al. (2004) built a general framework to help Arizona police departments investigate a wide range of crimes including but not limited to identity theft.
- Detect phishing attacks (Jakobsson and Myers, 2007) and identity theft that occurs through it.
- Detect credit card fraud, a subject closely related to identity theft, e.g., using the data mining approaches support vector machines and random forests (Bhattacharyya et al., 2011).

Such studies are, however, relatively few, possibly due to the lack of available data for research (Bhattacharyya et al., 2011). Our work, therefore, is novel in two ways:

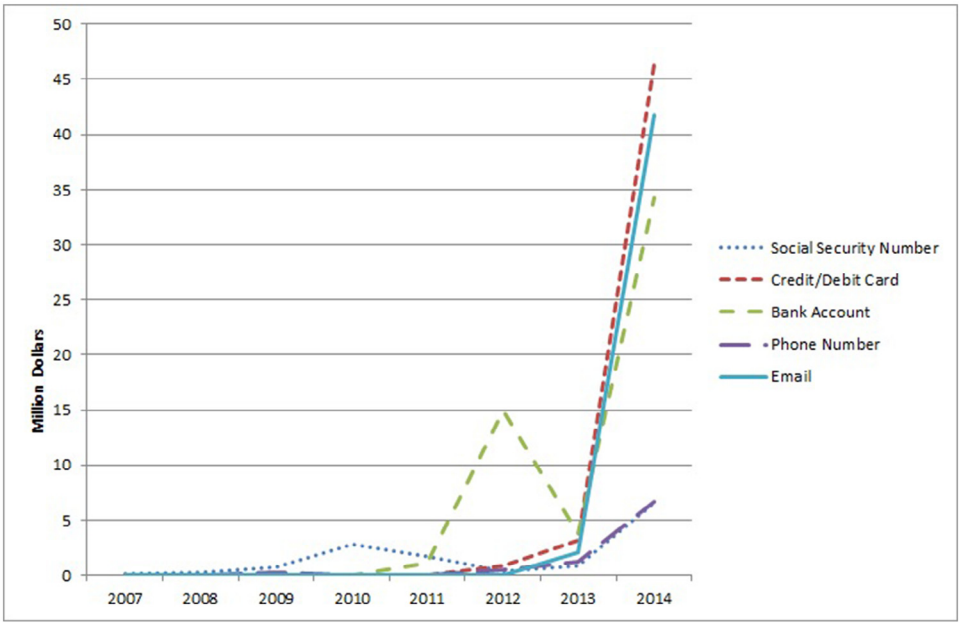

Fig. 11 – Loss per PII attribute per year (manual investigation).

- It takes advantage of text mining techniques to automatically extract information regarding identity theft and report statistics.
- It introduces news stories as a new source of raw data that can serve as a basis for future text mining techniques for identity theft.

8. Conclusion

This paper presents a proof of concept for collecting and analyzing publicly available news stories to assess and predict the behaviors of identity thieves and fraudsters. This research work used text mining techniques on news articles collected from the Internet to analyze and describe criminal behaviors and predict future trends of identity theft and fraud. The proposed process was automated except for the initial setup stage that involved predefining Personally Identifiable Information (PII) attributes and categories of behavioral criminal actions. The Identity Threat Assessment and Prediction (ITAP) algorithm was designed in a pipe-line manner where each step can be done separately and integrated together to build the entire analytical engine. The first step in the system involves the collection of news stories from the Internet. Around 3500 identity theft news stories were gathered. Subsequently, the text from these stories were preprocessed, eliminating irrelevant and unnecessary information. Then, named entities were extracted using the named entity recognizer. These named entities are categorized into different types, such as location, time, loss, etc., which together form an identity theft record. This record is then used to conduct analysis about different aspects of the identity theft, including the groups that experienced the identity theft, the risk of exposure for a particular PII attribute, the frequency of identity theft occurrence in different market sectors, the location of the identity theft, the potential financial impact caused by a compromised PII attribute, and the changes of such impact over time. Manual investigation of similar stories released in the same time period approves the overall trend of the results. Analysis of these results aim to increase empirical understanding of identity threat behaviors, offer early warning signs of identity theft, and thwart future identity theft crimes.

9. Future work

Mining identity theft news stories to better understand identity threat is very promising. Many aspects of this approach can be studied in more detail. Regarding the pipelined approach proposed in this paper, the following improvements and issues can be explored in future work.

First, the news media are prone to publish news that is "newsworthy". The identity theft stories with small amounts of loss may not be considered as newsworthy and these stories are less likely to be shown on the Internet. Therefore, the average loss for each incident calculated here may be higher than the true value. The influence of such bias on the news stories source may need to be taken into account. How to quantify such influence could be further investigated. Another issue is that the same identity theft could be reported in multiple news media. How to detect such duplicates and eliminate such influence is also worth studying. Future iterations of this model should seek to account for stories discussing the same event. Identification and combination of these stories may lead to increased nuance in the selection and processing of results.

Second, the approach to extract timing information from the news story could be improved. Currently, the timing information is obtained from the content of the news story. Could it be extracted from the HTML tags directly (may not be the time when the theft actual happens) or even better to build a hybrid model by combining the two approaches? The identity theft or fraud crime may cover several months and the story therefore has multiple dates. How the system interprets multiple timing data is also worth further studying. Should different date data be combined to generate an estimated date? A novel direction to improve the timing information is to build upon related work (Xu et al., 2016b,c) to automatically generate spatial and temporal relationships between concepts.

Third, this paper treats the frequency of a particular attribute's occurrence as the risk of exposure of this attribute. However, the results and analysis indicate that this may not be true for all the attributes. Certain attributes do not occur often in the identity theft news stories. For example, the attribute "zip code" does not have a high frequency of occurrence in the analysis. However since it is so widely used by people and easily obtained by the thief through various ways, it should have a high risk of exposure instead of a low risk. Therefore, quantifying the correlation between frequency and the risk of exposure and adding it to the calculation of risk may improve the accuracy of the result.

Finally, how to include new attributes to enrich the predefined attribute list is a difficult task. We investigated generating a potential attribute list for each news story. A potential attribute is a noun phrase identified by the POS tagger, which could be a new PII attribute not yet appearing in the existing list. However, new attributes still need to be manually selected from these potential attributes. Future research should consider the frequent words related to a single attribute as well as how the system deals with multi-word phrases.

Acknowledgments

We wish to thank the Center for Identity Partners (http://identity.utexas.edu/strategic-partners) for their contributions to this research effort.

REFERENCES

Aase K.-G., Text mining of news articles for stock price predictions, Norwegian University of Science and Technology; 2011.

Al A, Abu Q. IRS for computer character sequences filtration: a new software tool and algorithm to support the IRS at tokenization process. Int J Adv Comput Sci Appl 2013;4:81–2.

Allison SF, Schuck AM, Lersch KM. Exploring the crime of identity theft: prevalence, clearance rates, and victim/offender characteristics. J Crim Justice 2005;33(1):19–29.

Apache, 2016. Apache PDFBox [online].
Australia Bureau of Statistics, 2011. Personal fraud [online]..
AWARE Software Inc., Aware Software, Inc.: Awareness user guide. 2014.
Bhattacharyya S, Jha S, Tharakunnel K, Westland JC. Data mining for credit card fraud: a comparative study. Decis Support Syst 2011;50(3):602–13.
Boilerplate, 2016. Boilerpipe: Boilerplate removal and fulltext extraction from HTML pages [online].
Chen H, Chung W, Xu JJ, Wang G, Qin Y, Chau M. Crime data mining: a general framework and some examples. Comput 2004;37(4):50–6.
Choo K-KR. The cyber threat landscape: challenges and future research directions. Comput Secur 2011;30(8):719–31.
CIFAS, 2014. Depicting the UK's fraud landscape, CIFAS [online].
Copes H, Kerley KR, Huff R, Kane J. Differentiating identity theft: an exploratory study of victims using a national victimization survey. J Crim Justice 2010;38(5):1045–52.
CSID, 2016. CSID Identity Protection [online].
Dolan B, Quirk C, Brockett C. Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. In: Proceedings of the 20th international conference on computational linguistics. Association for Computational Linguistics. 2004. p. 350.
Erickson K, Howard PN. A case of mistaken identity? News accounts of hacker, consumer, and organizational responsibility for compromised digital records. J Comput Mediat Commun 2007;12(4):1229–47.
Federal Trade Commission, 2014. Consumer Sentinel Network Data Book for January–December 2014 [online].
Federal Trade Commission, 2016a. Consumer Sentinel Network Data Book, Federal Trade Commission [online].
Federal Trade Commission, 2016b. FTC Consumer Information [online].
Finkel JR, Grenager T, Manning C. Incorporating non-local information into information extraction systems by Gibbs sampling. In: Proceedings of the 43rd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics; 2005. p. 363–70.
Golden R. Newsferret: supporting identity risk identification and analysis through text mining of news stories [Master's thesis]. USA: The University of Texas at Austin; 2013.
Harrell E. Bureau of Justice Statistics, US Dept of Justice, Office of Justice Programs, Victims of identity theft, 2014. 2015.
Identity Theft Resource Center, 2014. Data breaches [online].
Jakobsson M, Myers S. Phishing and countermeasures: understanding the increasing problem of electronic identity theft. Hoboken, NJ: John Wiley & Sons; 2007.
Javelin Strategy and Research, Identity fraud report: card data breaches and inadequate consumer password habits fuel disturbing fraud trends. 2014.
Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the eighteenth international conference on machine learning, vol. 1. ICML; 2001. p. 282–9.
Lester M, Beason L. McGraw-Hill handbook of English grammar and usage: with 160 exercises. McGraw Hill Professional; 2012.
Lifelock, 2016. [online].
Manning C, Schütze H. Foundations of statistical natural language processing. Cambridge, MA: MIT Press.; 1999.
Marcus MP, Marcinkiewicz MA, Santorini B. Building a large annotated corpus of English: the Penn Treebank. Comput Ling 1993;19(2):313–30.
Morris RG. Identity thieves and levels of sophistication: findings from a national probability sample of American newspaper articles 1995–2005. Deviant Behav 2010;31(2):184–207.
Morris RG, Longmire DR. Media constructions of identity theft. J Crim Just Popu Cult 2008;15:76–93.
Newman G.R., McNally M.M., Identity theft literature review, United States Department of Justice: National Institute of Justice; 2005.
Pascual A., Miller S., Marchini K., Identity fraud: fraud hits an inflection point, Javelin Strategy; 2016.
Peng J, Choo K-KR, Ashman H. Bit-level n-gram based forensic authorship analysis on social media: identifying individuals from linguistic profiles. J Netw Comput Appl 2016;70:171–82.
Pratt TC, Holtfreter K, Reisig MD. Routine online activity and internet fraud targeting: extending the generality of routine activity theory. J Res Crime Delinq 2010;47(3):267–96.
Privacy Rights Clearinghouse, 2016. Privacy Rights Clearinghouse [online].
Quick D, Choo K-KR. Data reduction and data mining framework for digital forensic evidence: storage, intelligence, review and archive. Trends Iss Crim Crim Justice 2014; 480:1–11.
Quick D, Choo K-KR. Big forensic data reduction: digital forensic images and electronic evidence. Clust Comput 2016;19:723–40.
Reyns BW. Online routines and identity theft victimization further expanding routine activity theory beyond direct-contact offenses. J Res Crime Delinq 2013;50(2): 216–38.
Sanger J, Feldman R. The text mining handbook: advanced approaches in analysing unstructured data. New York: Cambridge University Press; 2006 ISBN 978-0-511-33507-5.
Serrano JI, Araujo L. Evolutionary algorithm for noun phrase detection in natural language processing. In: 2005 IEEE congress on evolutionary computation, vol. 1. IEEE; 2005. p. 640–7.
Teamshatter, 2016. [online].
The United States Department of Justice, 2016. Identity Theft [online].
Tjong EF, Sang K, De Meulder F. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In: Proceedings of the seventh conference on natural language learning at HLT-NAACL 2003, vol. 4. Association for Computational Linguistics; 2003. p. 142–7.
Toutanova K, Klein D, Manning CD, Singer Y. Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 conference of the North American chapter of the Association for Computational Linguistics on human language technology, vol. 1. Association for Computational Linguistics; 2003. p. 173–80.
Van der Meulen NS. Between awareness and ability: consumers and financial identity theft. Commun Strateg 2011;81:23–44.
Wang Y, Wang X-J. A new approach to feature selection in text classification. In: 2005 international conference on machine learning and cybernetics, vol. 6. IEEE; 2005. p. 3814–19.
Witten IH, Don KJ, Dewsnip M, Tablan V. Text mining in a digital library. Int J Digit Libr 2004;4(1):56–9.
Xu Z., Zhang H., Hu C., Mei L., Xuan J., Choo K.-K.R., et al., Building knowledge base of urban emergency events based on crowdsourcing of social media, Concurrency and Computation: Practice and Experience; 2016a.
Xu Z, Xuan J, Liu Y, Choo K-KR, Mei L, Hu C. Building spatial temporal relation graph of concepts pair using web repository. Inform Syst Front 2016b;1–10. doi:10.1007/s10796-016-9676-4.
Xu Z, Zhang H, Sugumaran V, Choo K-KR, Mei L, Zhu Y. Participatory sensing-based semantic and spatial analysis of urban emergency events using mobile social media. EURASIP J Wirel Commun Network 2016c;2016(1):44.
Yang Y, Manoharan M, Barber KS. Modelling and analysis of identity threat behaviors through text mining of identity theft stories. In: Intelligence and security informatics conference (JISIC), 2014. IEEE Joint; 2014. p. 184–91.

Razieh Nokhbeh Zaeem received her Ph.D. in Electrical and Computer Engineering from the University of Texas at Austin in May 2014. In 2010, she was honored as a Google Anita Borg Scholarship Finalist. She interned at Rockwell Automation Inc. in Austin, TX in 2010, and at Fujitsu Laboratories of America in Sunnyvale, CA in 2012. She has been a post-doctoral fellow with the Center for Identity since July 2014.

Monisha Manoharan received a Master's degree in Computer Science with a focus on Machine Learning and Data Analytics from The University of Texas at Austin, where she worked as a graduate research assistant in the Center for Identity. She currently works on data driven analytics projects at Schlumberger Technology Innovation Center.

Yongpeng Yang received his B.E. degree in Electrical Engineering from Harbin Institute of Technology in 2012 and M.S.E degree in Electrical and Computer Engineering from The University of Texas at Austin in 2014. Currently, he is a software Engineer at Google. Prior to joining Google, Yongpeng was a software engineer at Oracle, working on Oracle VM Server.

Suzanne Barber is the AT&T Endowed Professor in Engineering in the Department of Electrical and Computer Engineering and Director of the Center for Identity at The University of Texas at Austin. Previously serving as the Director of Software Engineering at The University of Texas at Austin, Dr. Barber led the cross-disciplinary Center for Excellence in Distributed Global Environments (EDGE).