The Thesis Committee for Yongpeng Yang
Certifies that this is the approved version of the following thesis:

Mining of Identity Theft Stories to Model and Assess Identity Threat Behaviors

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor:

Kathleen Suzanne Barber

Matthew McGlone

Mining of Identity Theft Stories to Model and Assess Identity Threat Behaviors

by

Yongpeng Yang, B.E.

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Engineering

The University of Texas at Austin

May 2014

Dedication

To my loving parents, Yang Baojie and Guo Yingzi, who are always by my side and support me.

Acknowledgements

I would like to thank my supervisor, Dr. Suzanne Barber, who patiently guided my research and has always been supportive of my endeavors during my stay here in The Center for Identity.

Thanks to Dr. Matthew McGlone, for consenting to be in the supervising committee, reviewing my work and providing valuable suggestions.

I would like to thank Monisha Manoharan, Liang Zhu, Muhammad Zubair Malik, Jeremy Golden, Elizabeth Goins and other lab members at the Center for Identity who suggested valuable tools and feedback during my research.

I thank all the professors, teaching assistants, staff, and administration at the Department of Electrical and Computer Engineering for their supportive and innovative work.

Special thanks to my dear friend Yichuan Niu, Ce Wei and Jennifer Brenner, who have always supported me through the graduate education.

Saving the best for the last, I would like to thank my mother and father for supporting me with the love, and care to help me get through the challenging process of completing a graduate degree.

Abstract

Mining of Identity Theft Stories to Model and Assess Identity Threat Behaviors

Yongpeng Yang, M.S.E.

The University of Texas at Austin, 2014

Supervisor: Kathleen Suzanne Barber

Identity theft is an ever-present and ever-growing issue in our society. Identity theft, fraud and abuse are present and growing in every market sector. The data available to describe how these identity crimes are conducted and the consequences for victims is often recorded in stories and reports by the news press, fraud examiners and law enforcement. To translate and analyze these stories in this very unstructured format, this thesis first discusses the collection of identity theft data automatically using text mining techniques from the online news stories and reports on the topic of identity theft. The collected data are used to enrich the ITAP (Identity Threat Assessment and Prediction) Project repository under development at the Center for Identity at The University of Texas. Moreover, this thesis shows the statistics of common behaviors and resources used by identity thieves and fraudsters — identity attributes used to identify people, resources employed to conduct the identity crime, and patterns of identity criminal behavior. Analysis of these results should help researchers to better understand identity threat behaviors, offer people early warning signs and thwart future identity theft crimes.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Identity theft is an ever present issue in our society, where almost all aspects of our lives are digital. According to the National Institute of Justice [1], "Identity theft has perhaps become the defining crime of the information age, with an estimated 9 million or more incidents each year." Over the past decade, the Federal Government and most states have passed legislation to impose criminal sanctions on identify thieves. Efforts to combat identity theft have been hampered, however, by the elusiveness of the definition and its overlap with the elements of many other crimes. Additionally, the long-term and multi-jurisdictional nature of identity theft, as well as the looming question as to whether law enforcement agencies or financial institutions are better equipped to combat it, add to the inability to fully contain the problem. Despite all stakeholders' awareness that there is a problem, it appears that no one is quite sure who should take ownership in solving it. Likewise, little time has been spent researching how identity theft actually occurs. There are best practices and prevention tips from security companies and government agencies available. What is void, however, is aggregated data about the process involved in stealing someone's identity. Most information available centers on reactive measures, which are helpful once your identity is stolen, but bring us no closer to ending or increasing the difficulty of future thefts. The consumers are typically several steps behind the identity theft.

To better understand the business process used by identity thieves and fraudsters, the Center for Identity at The University of Texas at Austin is developing a repository of relevant knowledge. The aim is to understand the criminal's business process, the vulnerabilities that allow the crime to take place, the resources that facilitate it and what can be done to prevent it. Armed with this knowledge, a shift in the definition and use of

credentials may be explored to decrease identity theft and fraud vulnerabilities. In order to better analyze the crimes that steal and use identity information, the Identity Threat Assessment and Prediction (ITAP, will be introduced later) tool is piecing together this business-like model of criminal methods and techniques. ITAP will allow us to better understand a fraudster's behaviors and inevitably, make connections and visualize patterns based on past identity theft and fraud. As more information is funneled into the tool, the ITAP will deliver actionable knowledge that is grounded in the study of thefts that have actually happened in the past. The big questions are: 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?

To assess and predict the identity threats, analytical tools like ITAP needs sufficient amount of data to conduct analysis and generate reliable results and conclusions. However, there is no publicly available repository describing ongoing identity theft and fraud, which would require many identity domain experts to derive and to subsequently generate well structured models of these criminal behaviors. Initially, the identity threat scenarios and data in the ITAP system of Center for Identity were entered manually. This data entry method is not sufficient to stay current with the quantity and pace of identity theft and fraud crimes. Thus, the biggest problem faced by the research team is that much of the identity theft data simply aren't available, even in semi-structured form, much less normalized, structured form. Although some government organizations have internal databases cataloging identity theft and fraud, almost none of them are available to the public or research institutions. The good news is that there are dozens of identity theft news stories published on the Internet each month. However, these articles are in a raw text format that cannot be analyzed directly to find patterns.

One way to collect this data requires many people reading those stories and then entering the useful information into the database manually. This process is very time consuming and, due to the magnitude of identity theft and fraud, it is difficult to remain current. To solve this problem, this thesis proposes an automatic solution that uses text mining, an application of natural language processing to extract the useful information from those identity theft stories and articles.

This thesis consists of six chapters including the introduction, conclusion and future work. Chapter 2 introduces some background information about the identity research domain and the ITAP project, as well as the commonly used text mining and natural language processing techniques. Chapter 3 talks about the design of the algorithms and the composition of the system. In Chapter 4, the results of the algorithm running on the news stories extracted online are shown and analyzed. Chapter 5 concludes with a summary of the thesis work. The possible future work is discussed in Chapter 6.

Chapter 2: Background

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

2.1 THE ITAP OVERVIEW

In order to concretely illustrate the ITAP model and its utility, we will first closely examine a specific form of identity theft – home equity fraud. This theft form involves multiple players and a series of well-articulated steps, as well as several resources and data elements.

Figure 1 displays a flowchart of a hypothetical home equity fraud scenario. In this scenario, the fraudster first became a loan officer in order to learn the inner workings of loan processing. This provided critical knowledge for him that he later utilized to perform the fraud. Leveraging this industry knowledge, he was able to collect mortgage information about wealthy couples and search for lease and loan documents in public databases. Next, he used a readily available graphics editing program called Photoshop was used to grab signatures from the loan documents. Typically, in order to carry out a fraud of any magnitude, a profile must be compiled of the intended victim. The rest of the victims' personal data were compiled via paid searches on skip-tracing sites, credit reports run on Experian, and ancestry.com. The fraudster then called the victims' bank with a resource called SpoofCard which allowed the projected number the financial institution viewed on their end, to be any number he chose. Cleverly, he chose the phone

number of the victim and used this to validate his identity when he called the financial institution. He requested wire transfer documents which he then applied the duplicated signatures to and faxed back. Lastly, he worked with several international partners to launder the money by sending it internationally and paying his partner to return it minus a transaction fee.

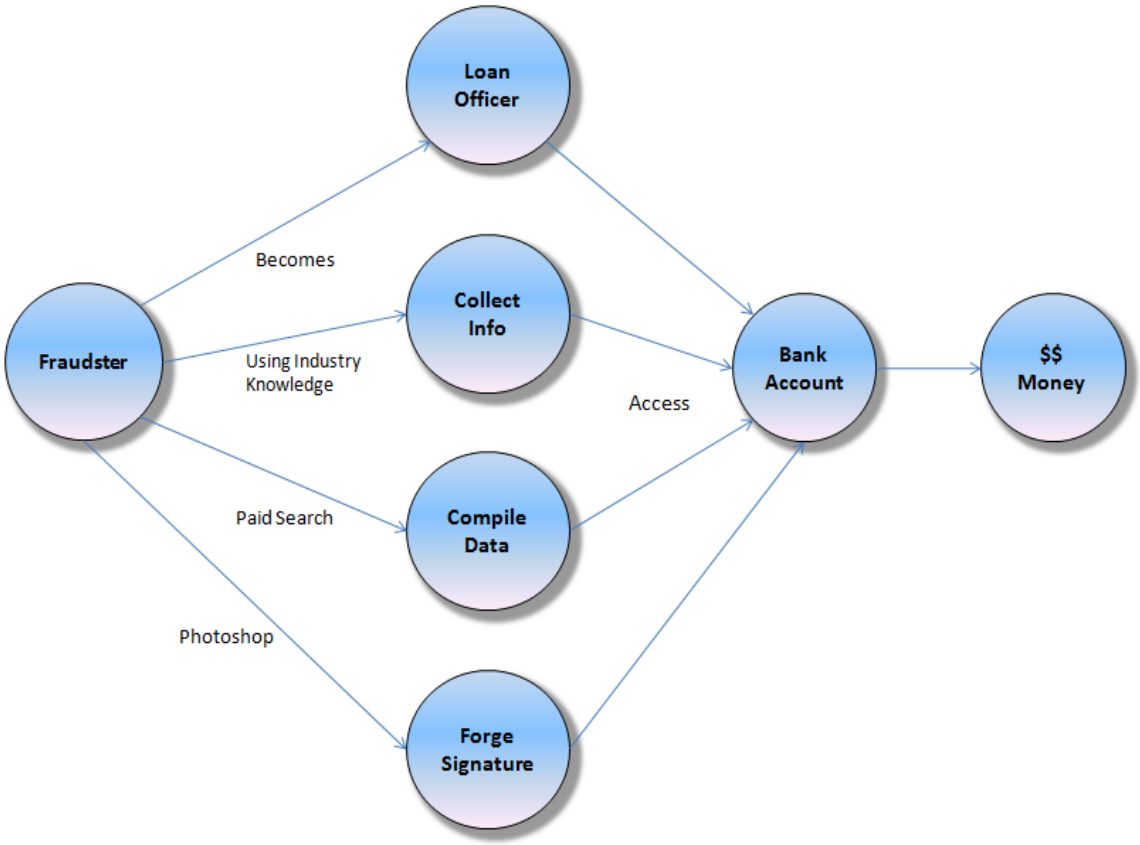

Figure 1 Home Equity Fraud Process

Perhaps the most unsettling part of this scenario is that the fraudster did not begin the fraud with a wallet, access to a bank account or a credit card number. He began with nothing. Through the manipulation of various vulnerabilities, online databases, knowledge of the inner workings of loans and financial institutions, and specific

resources, he was able to start with nothing and build profiles on victims until he eventually stole millions.

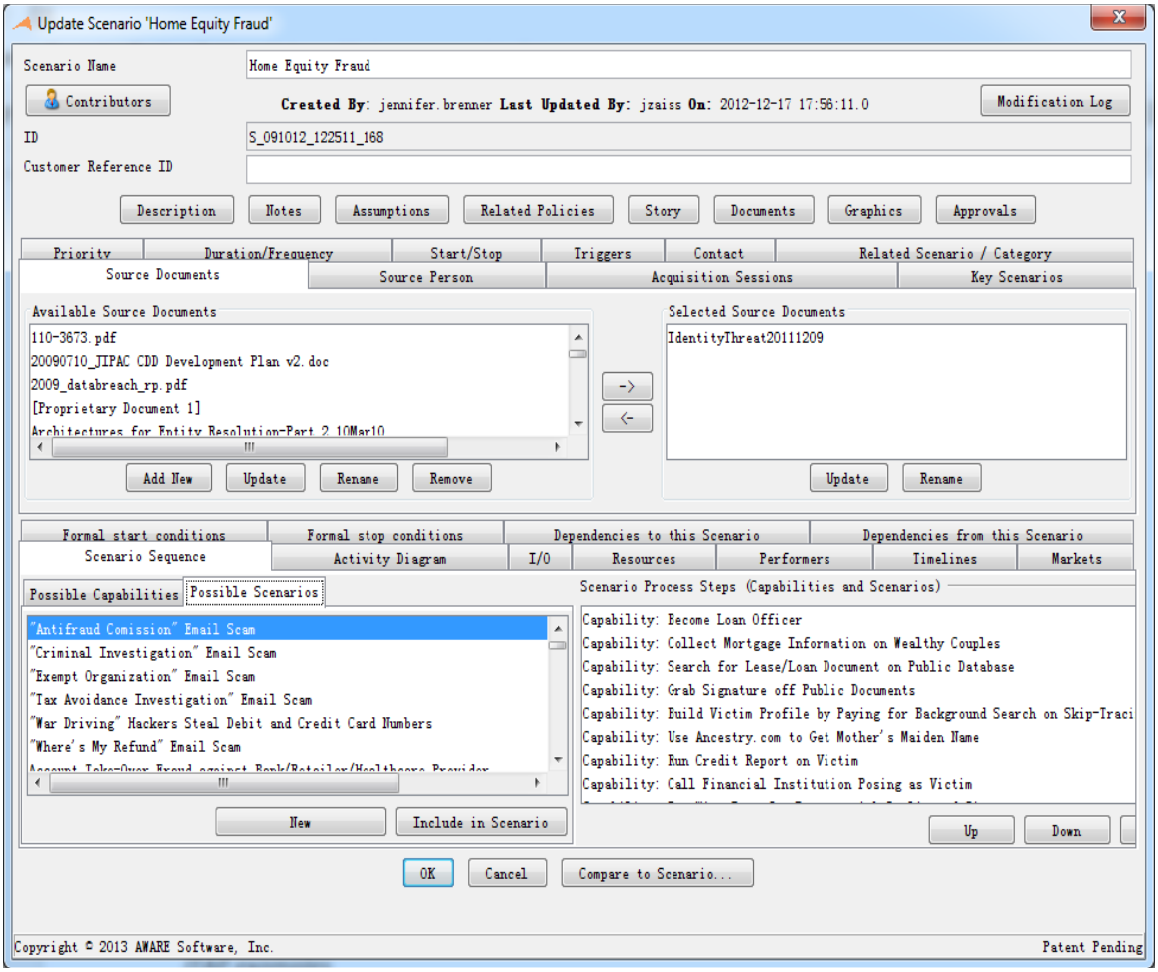

Figure 2 Home Equity Fraud Scenario [2]

Figure 2 depicts the Home Equity Fraud scenario in the ITAP view. The capabilities, or steps, are listed in the order they took place during the course of the fraud. This helps shed light on what progression of steps were necessary for the fraudster to carry out the fraud as a whole.

2.2 ITAP MODEL

This section describes the ITAP model representation and each component, and then articulates how the data in each scenario are stowed in the ITAP for further analysis.

2.2.1 Identity Theft as a Business Process

A business process is defined as a collection of related, structured activities or tasks that produce a specific service or product for a particular customer or customers. It often can be visualized with a flowchart as a sequence of activities with interleaving decision points or with a process matrix as a sequence of activities with relevance rules based on data in the process [3]. The process of committing identity theft mirrors a typical business process where each step serves a particular goal in the overall theft. Resources as well as input and output data elements allow the fraudster to advance from one step in the process to the next. As with any business process, if a critical step is missing or cannot be completed, the business process as a whole is halted. By viewing identity theft as a business process, ITAP provides a better insight and understanding of how the identity thieves conduct the crime step by step. This enables us to find the most critical part of the whole process, i.e. the most vulnerable step, and come up the countermeasures to prevent it. Like any other business process, without all of the necessary components, the process cannot be carried out. Below, each piece of the ITAP model is discussed in detail, and in conjunction, create the identity theft business process.

2.2.2 Model Representation

The ITAP model consists of several components, which are used to describe the different parts of the whole identity theft process and to analyze it. These components

together represent the process of an identity theft, including the steps taken by the thieves, the tools and resources they use and the capabilities of the thieves. It also provides possible solutions to prevent the theft from happening. By carefully examining each component, one can better understand the whole process of the identity theft and strengthen the weakest point in the process. Figure 3 gives an overview of the ITAP model. Each component in the model will be discussed in detail later.

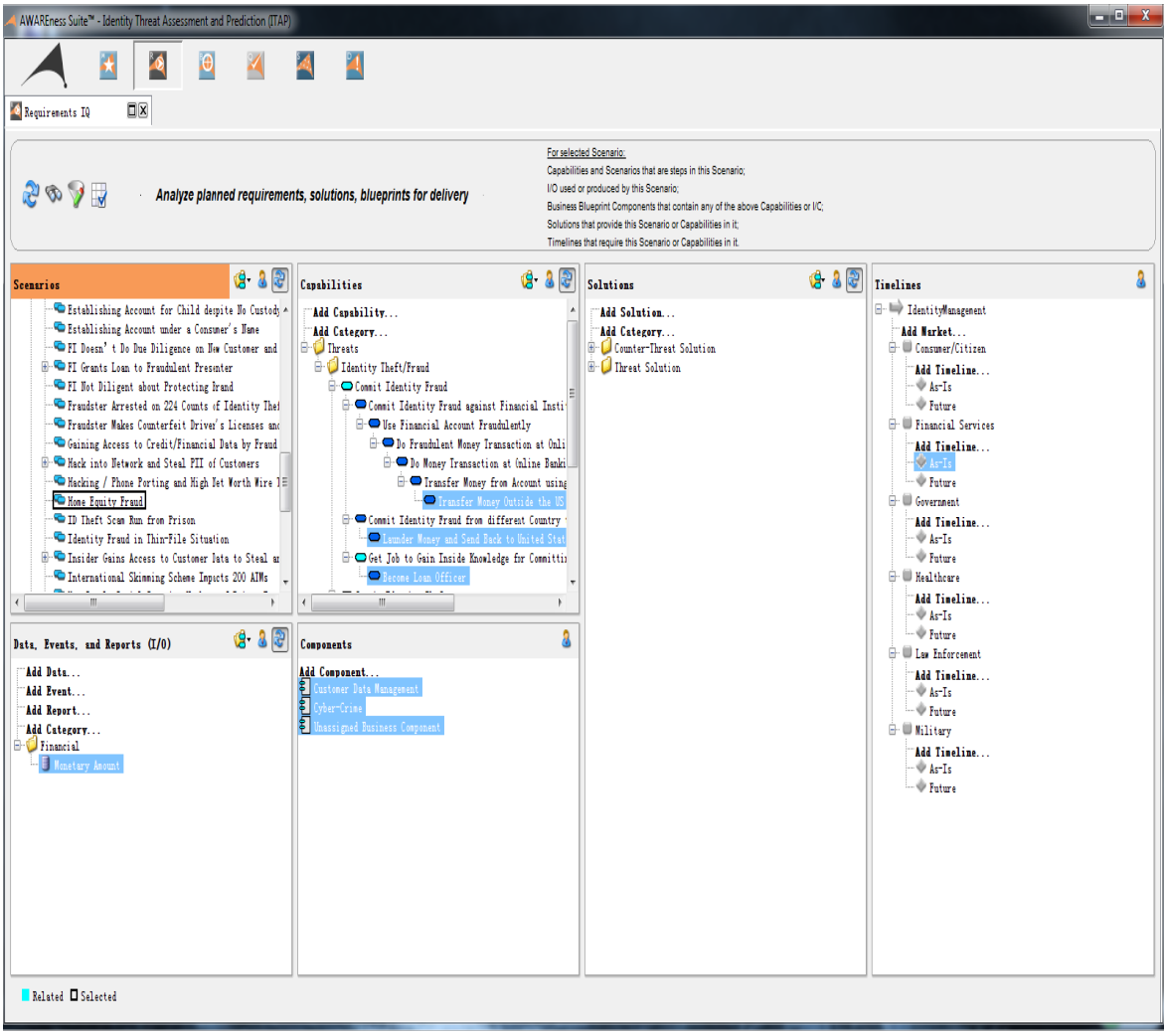

Figure 3 ITAP Model Overview [2]

2.2.2.1 Scenarios

Figure 4 depicts a snapshot of the scenarios, or stories, currently in the ITAP. A scenario refers to a description of the interaction sequences between the human user and the automated system. A scenario consists of several steps, which can be a function, a sub-sequence or an event. A sub-sequence, which is also known as a sub-scenario, is a particular sequence of the functions within the given scenario [3]. The steps are displayed in the occurring order which is beneficial for the analyst to reason about the cause and effect of the theft. There could also be some overlapping sub-sequences between different scenarios. A sub-sequence occurring in many scenarios indicates that this sub-sequence is used by the identity thieves to commit crimes as common steps. Identifying such sub-sequences can help us find the commonalities among the identity thefts and thus come up with general solutions to deal with these thefts.

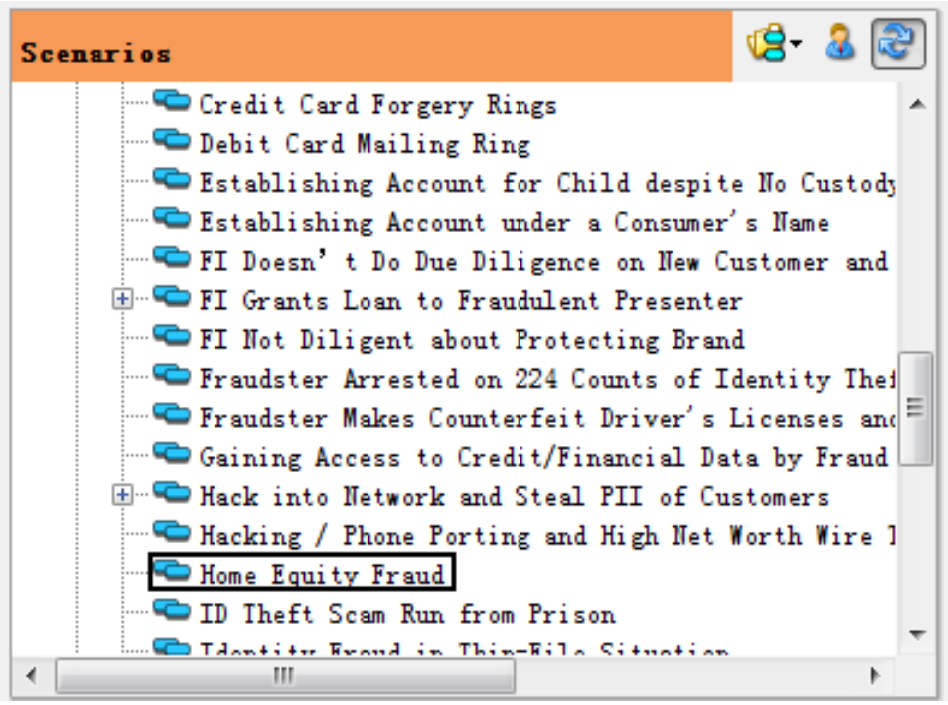

Figure 4 ITAP Scenarios [2]

2.2.2.2 Inputs and Outputs

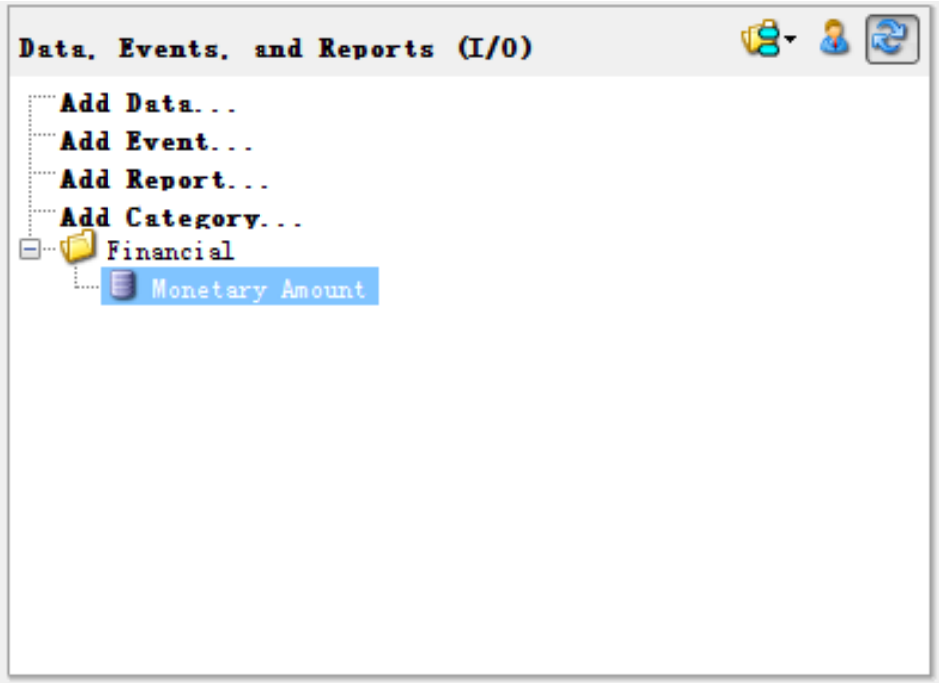

Figure 5 ITAP Inputs and Outputs [2]

Figure 5 depicts a snapshot of the inputs and outputs (data, events, and reports), which refer to the information that is transferred between different scenarios, functions, systems and their users. A data element is a token that represents useful information and can be interpreted as some kind of value and it can be read or written/modified for certain purposes. An event is an action or behavior that occurred and is detected by the system, which may needs further handling and can affect the system state. A report refers to the information created by the system for human comprehension with or without further interpretation. A data item could be of an integer, an array, a string or some more complicated container. An event can be a keyboard input or an exception, etc. A report could just be an email or a table that generated by the system [2].

2.2.2.3 Capabilities

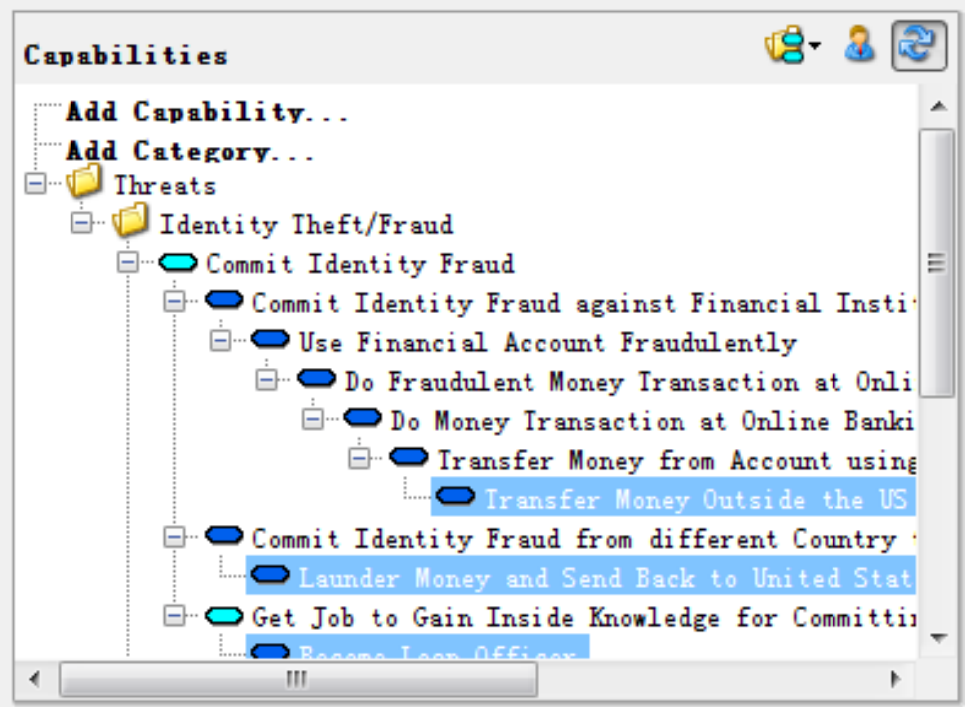

Figure 6 ITAP Capabilities [2]

Figure 6 depicts a snapshot of the capabilities component, which describes the actual steps, or things the fraudster had to have been capable of doing, in order to achieve the overall fraud. For example, the ITAP currently describes a very elaborate scenario of home equity fraud. The first step, or capability, includes the fraudster obtaining a job as a Loan Officer to learn the internal processes involved in processing loans and other such related documentation. This is considered as a crucial step in the overall scheme since the knowledge he acquired here inevitably allowed him to pull off the entire fraud. The thief was able to learn exactly what the proper procedures were in handling loan documentation, how authentications were handled by financial institutions when clients were calling to check on statuses as well as which types of banks were the easiest targets.

2.2.2.4 Components

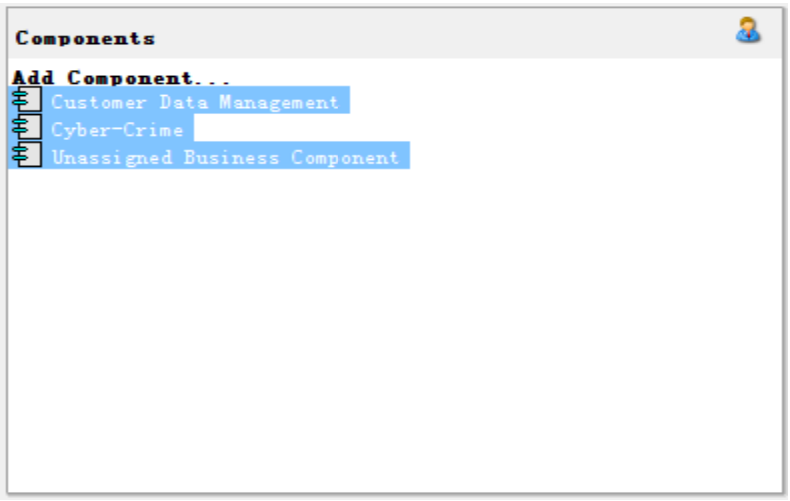

Figure 7 ITAP Components [2]

Figure 7 depicts a snapshot of the component, which refers to an entity that is able to perform certain systems functions and its associated inputs and outputs. A component consists of functions, data, reports, and events.

2.2.2.5 Resources

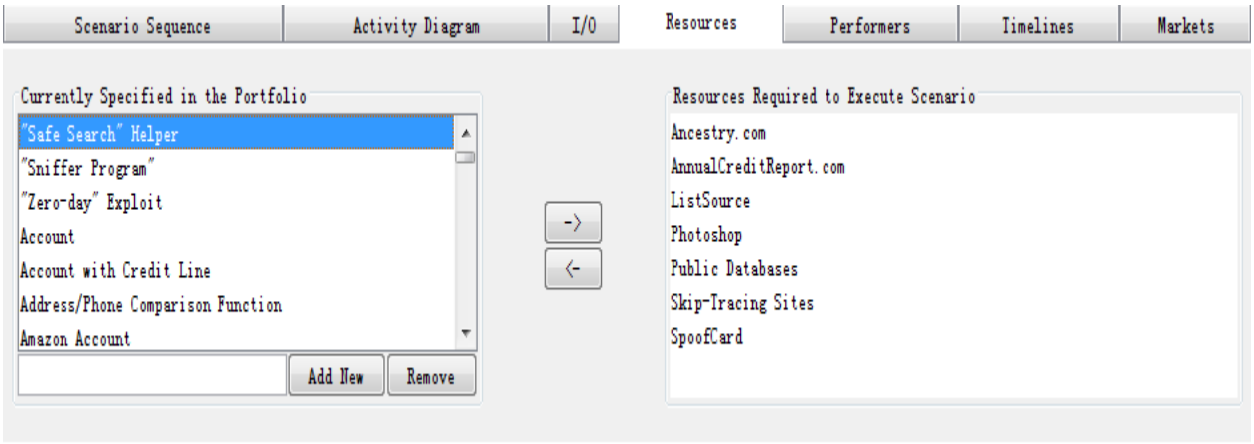

Figure 8 ITAP Resources [2]

Figure 8 describes the resources that are used by the fraudster to complete a step as well as the overall fraud. A resource can be anything from malware code, a call spoofer or a credit card skimmer to easily accessible software, such as Photoshop. Anything that the fraudster physically uses can be considered a resource, and unfortunately, our research indicates that many of these resources are readily available.

2.2.2.6 Performers

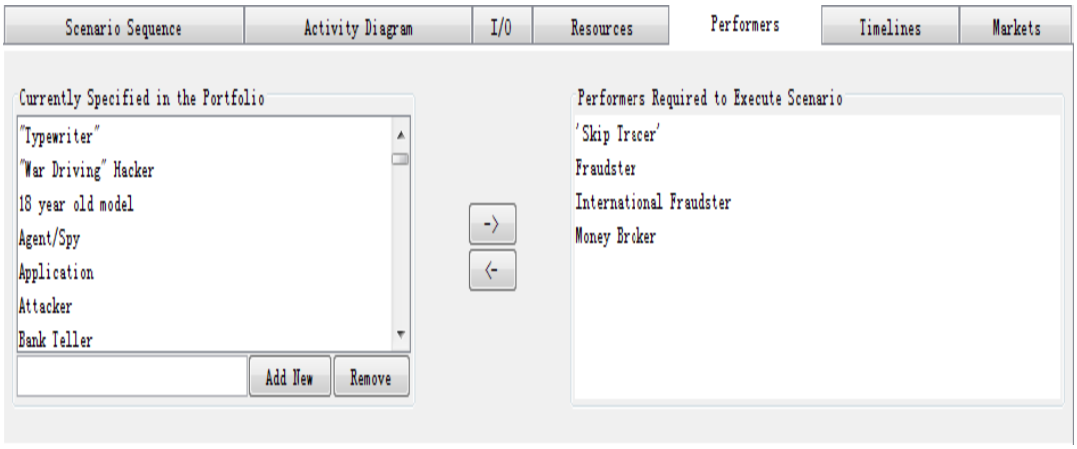

Figure 9 ITAP Performers [2]

Figure 9 presents the performers who played a part in executing a scenario or completing the step. It could be a hacker, a skimmer or even someone that internationally plays a part. It is very important to assess who the key players are in each step of the overall fraud. In doing so, one is able to view the big picture of how many and what types of people were necessary in carrying out the attack. In the home equity example, although there was one major fraudster, the thief did receive assistance internationally. He worked with an international fraudster and broker when transferring the stolen money outside the US. This allowed him to send the money outside the US and

bring it back inside in an attempt to avoid being caught. With the assistance of these two additional people, he was able to launder roughly $7 million dollars every two weeks until he was eventually caught.

2.2.2.7 Timelines and Market Segments

Scenarios are categorized into different market segments, which allow us to understand where threats are taking place and what industries may be most susceptible to attack. Each threat is categorized as either being "as-is" or "future". "As-is" describes a threat that can or has happened in the current environment within that market segment. "Future" highlights threats that may be seen in a future environment. Figure 10 depicts the timelines and the market segments.

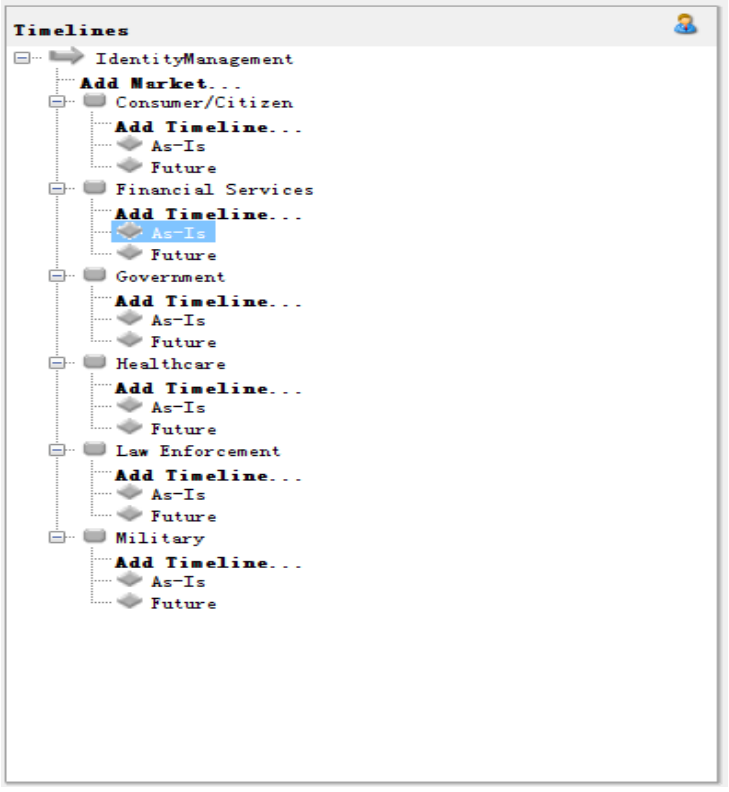

Figure 10 ITAP Timelines and Market Segments [2]

2.2.2.8 START and STOP Conditions

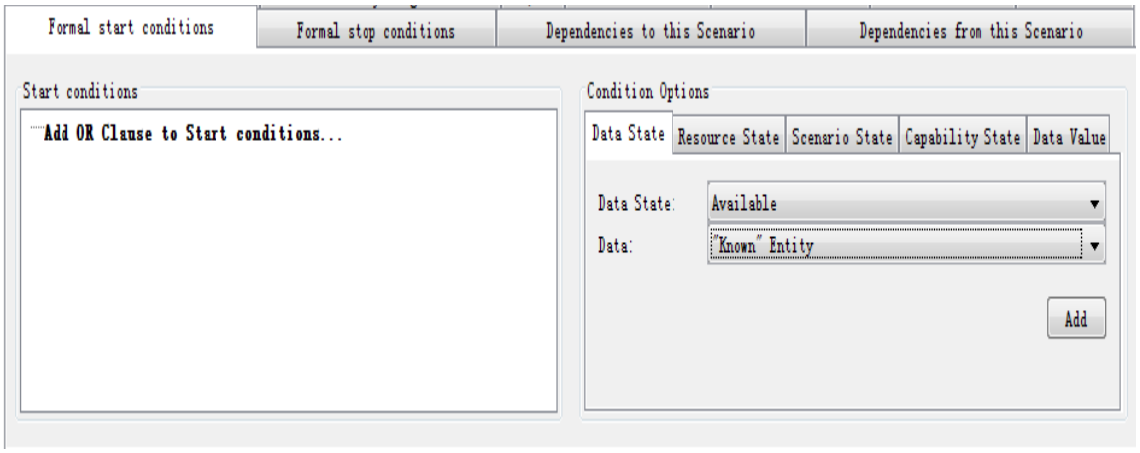

Figure 11 ITAP START and STOP Conditions [2]

Another important feature of the ITAP is the ability to link specific steps together by adding START and STOP conditions, which are shown in Figure 11. What this does is to allow us to link steps together in the event that one step is absolutely required to be completed before the next step can begin. The steps are linked together, if and only if, one would always follow the other. This allows us to create true connections between the the steps themselves, which upon further analysis can throw red flags when similar pairs of steps occur in the future that identity fraud may be looming. For example, in our Home Equity Fraud scenario, the fraudster was able to run a credit report on the victim on annualcreditreport.com with a goal of gaining specific HELOC details. The HELOC details, or Home Equity Line of Credit details, provide information on a specific type of loan the fraudster used to carry out his scam. Because the data input to running the credit report required knowledge of the victim's address, date of birth and social security number, the preceding step, the building of the victim's profile by paying for a background search on a skip-tracing site, is a necessary START condition. That is, the

fraudster needed to complete the background search prior to being able to access the credit report. Now clear connections can be made on how the information is flowing and what the dependencies look like between certain steps. This type of pattern detection will will prove invaluable in predicting future identity theft scenarios.

2.3 TEXT MINING

Text mining usually refers the process of gleaning the meaningful information from natural language text. The goal is to analyze the text and extract the useful information for a specific use [4]. It is essentially an application of natural language processing to transform the natural text into directly usable data. 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 finally evaluating and analyzing the generated data. Typical text mining research includes text clustering, text categorization, entity extraction, sentiment analysis, entity relation modeling and so on. Text mining techniques have been used in many areas to help process large amount unstructured data, such as biomedical applications (e.g. association of gene clusters and identification of biological entities), 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) [5].

This section briefly introduces some text mining techniques used by common text mining systems, especially those methods used or planned for use in this thesis. Some methods not used in this thesis but used by other text mining applications are also described here.

2.3.1 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 [6]. Tokenization is one of the most common techniques used for text preprocessing.

Tokenization 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 [7]. There are plenty of ways to tokenize a text stream into meaningful tokens. One simple approach would be just split the text or sentences based on the white spaces, punctuation, or other special symbols between words.

After tokenization, stop-word removal and stemming (lemmatization) may apply for further preprocessing [8]. Stop-words refer to high frequency words in a language that don’t 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 to reduce noise and to select meaningful textual features.

Stemming (lemmatization) is the process of reducing inflected words into a stem or base form so that the number of phrases or words with similar meaning can be 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 same stem ‘look’. It is usually beneficial to map all inflected forms into the stem. However, some experimental results show that sometimes stemming can have a negative effect on a text classifier [9]. The stemming process can become complicated for

some words or phrases which have many exceptional cases, such as ‘be’ and ‘was’, ‘see’ and ‘saw’. In the identity research area, words such as ‘social security number’, ‘SSN’, and ‘social security card’ refer to the same identity attribute. Thus, by combining these words into the stem, ‘social security number’, will reduce the complexity. The most commonly used stemmer is the Porter Stemmer, which transforms a word into its base form based on a set of language specific rules [10].

2.3.2 Features Extraction

In addition to text preprocessing, feature extraction is also an important step before the natural language text can be analyzed by text mining techniques. The text document is often too large, as well as redundant, to be processed by some mining algorithms, thus a set of features is produced to represent the text document that reduces its dimensionality [11]. This process is called feature extraction. The features yielded by this process are also referred as feature vectors.

Feature vectors could be primarily lexical and character features as well as other semantic or higher-level features [12]. Primarily lexical and character features are the most widely used ones, which are word–based features that can be observed directly in the text, such as word frequencies, n-grams, noun phrase, vocabulary richness, and character n-gram, etc [13]. The major advantage of low-level features is that they can be extracted easily in an automatic fashion. Also, these features are easier for humans to understand and reason about. On the contrary, semantic features or higher-level features are extracted from the surface level lexical features by using statistical techniques, such as singular value decomposition (SVD), topic modeling, and random projection, etc. The

higher-level features can capture more semantic information in a text document and thus are great for performing tasks like classification, clustering and so on.

2.3.2.1 N-Grams

An N-Gram is a subsequence of n items from a given sequence. In the context of text mining, a word constitutes an item. Two commonly used N-grams, unigrams and bigrams, are illustrated as examples below.

Unigrams are N-Grams of size one, i.e. one single word. They are usually made of all the single words that consist of the text document after preprocessing. The set of unigram features are also known as the “bag of words” feature sets. Although the unigram model is quite simple, it has achieved success in text classification and word sense disambiguation [14]. Bigrams are N-Grams of size two, which are a consecutive sequence of two words, and are usually used as the basis for simple statistical analysis of text. Bigrams captures more underlying information of the text structure than unigrams, which might be beneficial for tasks like text classification and clustering.

For example, the unigrams generated from two sequences “He likes basketball and hates football.” and “He likes football and hates basketball” are identical. However, by using bigrams, “likes basketball” and “hates football” generated from the first sequence express completely contrary meaning from “hates basketball” and “likes football” generated from second sequence. Thus these features can be used to distinguish two sequences.

2.3.2.2 Noun Phrases

Noun phrases (NP) refers to units whose first or principal word is a noun, pronoun or other noun –like words, which can be modified by words such as adjectives [15]. 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 to the nouns that represent unique entities [16], such as San Francisco, LeBron James, or Miami Heat. 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.

2.3.2.3 Singular Value Decomposition

Singular Value Decomposition is a matrix factorization approach that has been used in many applications. The key idea of SVD is to replace the original feature-document matrix with a newly generated but much smaller matrix. The features in the new matrix represent the latent features and maintain the characteristics of the original features approximately. The latent feature represents certain properties of the objects that have not been observed directly, or represents the hidden causes that could explain the observed properties [17]. For example, several features of an object may always show up together. Instead of using all of them, using a latent feature to represent this characteristic will reducing the complexity. Thus SVD is also regarded as a feature reduction. It can be used to reduce the dimension of a complicated file and represent it in a simpler way without losing the essential information.

In text mining, SVD has been used in LSA (Latent Semantic Analysis), which is a well known technique to map the high-dimensional count vectors, such as the ones occurring in vector space representations of the text documents, to a latent semantic space, which is a reduced dimensional representation[18].

2.3.3 Document Representation

Natural text documents are usually too large and hard to deal with. Therefore the documents are often transformed to N-dimensional vector. Each dimension represents a characteristic of the document. The characteristic could be a feature extracted from the text as described earlier, such as words, phrases or some other elements, which are weighted according to the importance [19].

2.3.3.1 Vector space model

A common way of representing the document in text mining is to use a vector space model, which uses a multi-dimensional vector. Each dimension exhibits a feature extracted from the natural text [20]. The vector space model is typically used to compare the similarity between two document vectors. This comparison is made by calculating the cosine of the angles between two vectors. It can also be used to answer queries about the text documents. The similarity can be calculated using the formula below:

$$similarity(D_1, D_2) = \cos\theta = \frac{D_1 \cdot D_2}{\|D_1\|\|D_2\|}$$

Where D_1 and D_2 are two vectors that represent two documents. The cosine is the normalized for product of the two vectors. A zero similarity means that the two documents do not share any common features in this vector space because the angle between the two vectors is 90 degrees (orthogonal).

2.3.3.2 Term Weighting ($tf - idf$)

In vector space, different features are assigned different weights. A higher weight of a feature means that it has greater impact on the cosine similarity. Thus the more important feature should be given a higher weight. How to decide whether a term

(feature) is important or not? The three most commonly used major factors that affect the importance of a term are the term frequency factor (tf), inverse document frequency factor (idf), and document length normalization factor [21]. The term frequency factor refers to the frequency with which a term occurring in a document. The simplest calculation is just counting the frequency of a term in a document by using the formula:

$$tf_{t,d} = f_{t,d}$$

Where $f_{t,d}$ is the frequency of a term t occurring in a document d.

Inverse document frequency measures a term's scarcity across the document collection. It can be calculated by dividing the total number of documents by the number of documents containing that term, and then taking the logarithm of the quotient. The formula to compute the idf is:

$$idf_t = \log\left(\frac{N}{n_t}\right)$$

In this formula, N is the total number of documents in the collection and n_t is the number number of documents that contains term t.

The document length normalization factor normalizes the effect of document length on the document ranking by adjusting the term frequency or the relevance score.

The $tf - idf$ weight (term frequency–inverse document frequency) is the most commonly used term weighting method in text mining. It measures the relative frequency of a given term in a particular document compared to the inverse proportion of the term over the entire document corpus. In other words, it calculates how relevant a given term is in a specific document [22]. $tf - idf$ weighting can be calculated using the the formula below:

$$(tf - idf)_{t,d} = tf_{t,d} \times idf_t$$

2.3.4 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 (IE), which locates and classifies the words or phrases into predefined categories [23]. NER systems have been implemented by using a variety of models, such as Hidden Markov models (HMMs), Maximum Entropy Markov models (MEMMs) and Conditional Random Fields (CRF) [24]. Stanford’s NLP research 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 [25].

2.3.5 Part-Of-Speech Tagging

Part-Of-Speech Tagging (POS Tagging) is a process that assigns a word or a phrase in a corpus (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 could 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. POS taggers have been developed by many research groups using various models. The most common two approaches 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 corpora such as the Penn Treebank [26]. The learning-based approach has proven to be more effective considering the devoted human effort and expertise. Stanford POS tagger used in this thesis is built by the Stanford NLP group, which combines multiple features with a Cyclic Dependency Network that has 97.24% accuracy on the Penn TreebankWSJ, reducing the error by 4.4% compared to the best previous single automatically learned tagging result [27].

2.3.6 Typed Dependency

There are two common ways to represent the structure of sentences. The first approach is using phrase structure parsing, which is based on the constituency relation and represents the sentence structure as nested constituents. In contrast, the other method, known as typed dependency parsing, represents the dependencies between individual words. Typed dependency parsing also describes the grammatical relations between different words, such as subject or indirect object [28].

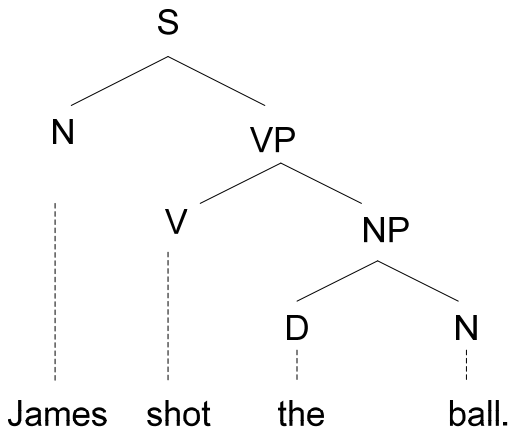

Figure 12 Phrase Structure Parse Tree

For the simple sentence “James shot the ball”, a phrase structure parse tree is shown as Figure 12. S, sentence, is the top-level structure in this example. N stands for noun. The leftmost N, “James”, is the subject of the sentence. The second one is the object of the sentence. VP, verb phrase, serves as the predicate. V, verb, is a transitive verb “shot” here. NP, noun phrase, is “the ball” here. D, determiner, is the definite article “the” in this example.

A typed dependency parse tree of the same sentence is shown as Figure 13. This parse tree does not have the phrasal categories (S, VP, and NP) seen in the Phrase structure parse tree above. The other notations are the same as the ones described previously.

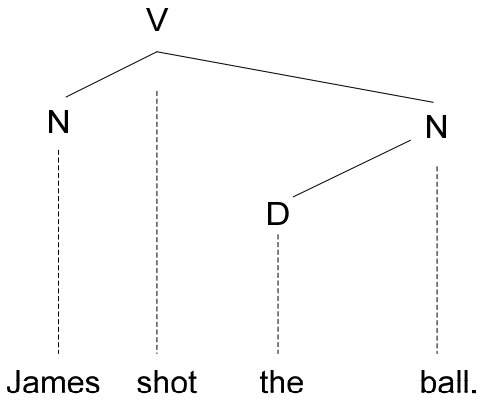

Figure 13 Type Dependency Parse Tree

Chapter 3: Algorithms and Design

The previous chapter briefly described the ITAP project as well as some preparatory work and common techniques used in text mining. This chapter will explain the algorithms and design that are used to mine the news articles/stories gathered from the Internet. The idea is to design a pipelined 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.

3.1 HYPOTHESIS

The hypothesis follows. Online 'identity theft' related news stories represent a reasonable sampling and description of the identity theft, including when, where and how it happens, the resources involved in the theft, and the loss caused by the theft to some extent. The ideal circumstance is that the accurate identity theft report provided by the investigation agency is available. However, even the investigation agency may not have the accurate information due to victim's obliviousness or lack of knowledge regarding the criminals' processes. By representing each news story using an identity theft record, the identity theft can be evaluated in a more detailed way and the analysis can reveal insights about the crimes and criminals. An identity theft record in this thesis refers to a predefined representation of an identity theft or fraud crime in the ITAP.

The data for some entries in the identity theft record representation are missing due to the nature of the news story. Thus this representation is not a complete one to reflect an actual identity theft and fraud. However, as mentioned before, even the original source document gathered by the investigation agency is not complete. Also, the news media tends to report stories that are considered as 'newsworthy' [29]. Quantifying such

bias and its influence on the data set generated from the news articles is difficult. For example, an identity theft resulting in a small monetary loss may not be considered as 'newsworthy' so this data is not reported and not included when calculating the averaged loss for each incident. Thus, the average calculated losses maybe higher than the ground truth value. Possibilities for evaluating the influence of such bias in the future are mentioned in the last chapter.

Although there are some limitations, news stories have several important characteristics that are beneficial in the ITAP analysis:

1. There is a tremendous amount of identity theft news stories. The news media publishes large numbers of news stories every day.
2. They are publicly available. One of the problems in the identity research area is that it is difficult to obtain well-formatted source data from the government or corporations. However, the information published in a news story is publicly available. Therefore, researchers do not need to worry about protecting the Personally Identifiable Information attributes associated with reported victims.
3. Mostly reliable. Although some information published in the Internet is not accurate or even false, most news stories are reliable and trustworthy since the news media is responsible to the public for providing accurate information.

3.2 PIPELINED SYSTEM MODEL

The designed pipelined system model is shown in Figure 14. The system first obtains the news stories from the Internet. Then the story text is preprocessed and irrelevant and unnecessary information is eliminated. After that, the named entities are

extracted by using the named entity recognizer. These named entities are then 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. At the same time, the system analyzes the typed dependency for each sentence in the story. The typed dependency is then used to generate the sequence diagram.

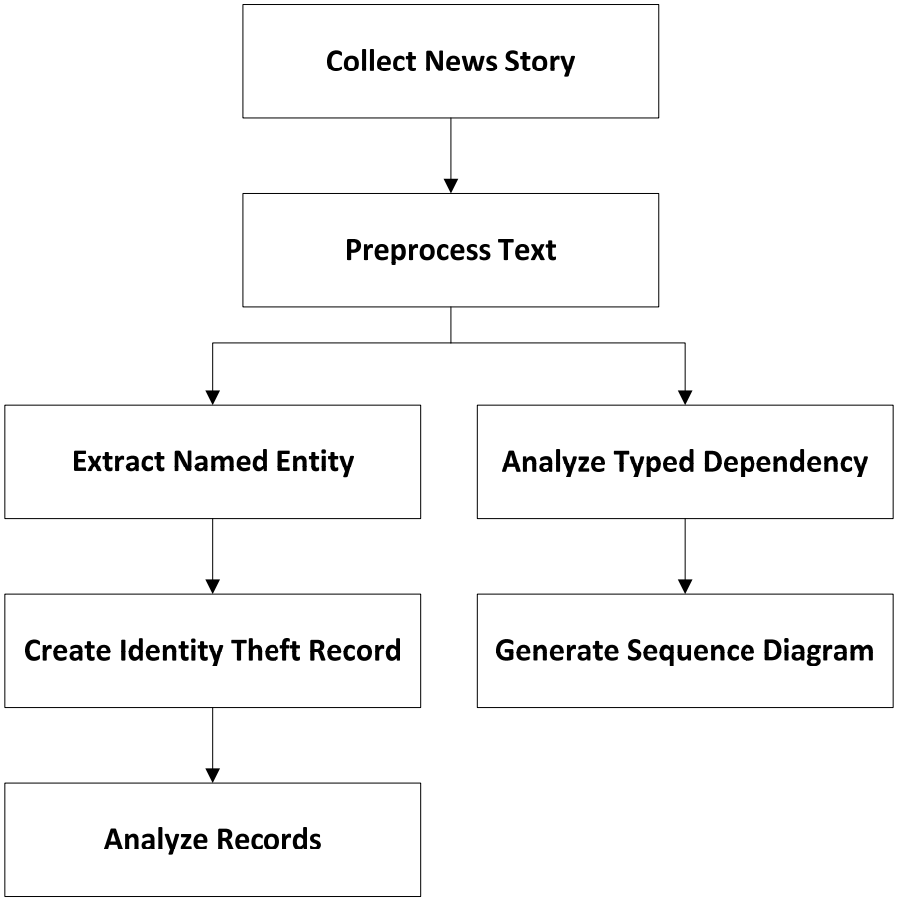

Figure 14 Pipelined System Model

3.3 NEWS ARTICLES COLLECTION

The first step in the data collection process is to get the news articles' links from the Internet. The links are gathered using Google search engine and other news feeds to

search a set of key words that are highly related to ‘identity theft.’ Some links are also extracted from the publicly available annual identity theft reports [30]. The next step is to retrieve the news articles based on the previous links. The main text content of the articles is exported to a single text file where the clutters around the main content in a web page are removed.

To extract news articles from the published format(usually HTML), the boilerpipe library is used to obtain the main content of the news stories and remove irrelevant content, such as format labels, navigation lists, advertisements and so on [31]. Moreover, sometimes the links points to PDFs that contain the article instead of a HTML file. To address this problem, a PDF extractor, developed based on the PDFBOX library [32], is used in such circumstance to extract the stories and stores the content in a text file. Links pointing to an invalid URL address are just simply discarded.

In order to keep track where the source comes from, the original links to the articles are stored along with the stories.

3.4 DEFINE PII ATTRIBUTE

A pre-existing list of Personally Identifiable Information (PII) attributes is defined manually by selecting the commonly used identity attributes by the identity thieves. This list will be enriched during the ITAP data collection and analysis as more and more attributes appear in the new stories. These attributes will be used to build the “bag of words” model [20]. The words in the pre-existing attributes list are compared against the words in the news articles and the matched words will be stored into the corresponding identity theft story record. For example, the attribute ‘social security number’ is a predefined PII attribute. The system will look up ‘social security number’ in the articles

when processing the new stories and count the frequency of occurrence of it. Such information will be stored and used later to conduct further analysis.

3.5 TEXT MINING

This section will describe the ITAP algorithms and approach used to mine the news stories gathered from the Internet. After processing, each story is represented by a DAT file, which stores all the useful information related to the original story. Each DAT file is essentially a java HashMap object, which consists of the "Victim", "Organization", "Location", "Date", "Cost", "Resource", "Actions", "SourceLink" and so on. This DAT file servers as an identity theft record and will be used as the data input to the statistical program.

3.5.1 Article Text Processing

Before actually processing the news stories, “preprocessing” is first done to eliminate some unused language-depend factors (like space in English, some language doesn’t have space). The preprocessing step used in this thesis is tokenization, which is implemented based on PTBTokenizerAnnotator from the Stanford CoreNLP library [33]. The PTBTokenizerAnnotator is a PTB (Penn Treebank) style tokenizer. The news article is tokenized and each token’s character offset in the article is saved. The Stemming (lemmatization) will be done when analyzing the data and will be described later.

The news article is then parsed by again using the Stanford CoreNLP library which enables the named entity recognition (NER) and part-of-speech (POS) tagger function. Each named entity (words or phrases) in the news stories is assigned to a corresponding category, such as people, organization, money, time and so on. These

named entities are candidates for the features to represent this news story. The next step is to analyze these named entities and determine if it is a valid attribute and store the valid ones into the DAT file. The POS tagging for the words will be used later in the identity theft sequence generation.

3.5.2 Time Selection

The time of occurrence for an identity theft is important to produce the correct analysis, such as calculating the loss and risk changes related to a timeline. Due to the nondeterministic duration of the identity theft and the delay of publishing time of the news story, it is hard to get an accurate time of the identity theft. There are two approaches to analyze approximate time from the news articles.

The first method is to choose the time that the articles are published. The problem with this approach is that the identity thefts might have happened long before the news articles are actually published. And in most cases, the news is published at least some time after the theft happened since getting the incident record or interviewing the victims are usually several days or weeks or even months later. So this method could extract a time that does not accurately reflect when the identity theft really happens.

The second method is to choose 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 contents of each article and highly related to the theft itself. This works in most cases, but some special conditions need to 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 will weigh equally when used later for further analysis. When an article doesn’t even

have a valid time in the main story content, should the date information be marked unavailable or just use the date the article was published instead? If the latter one is used, will that cause an inconsistency issue? And the tags in the HTML file for storing the date are quite different from each other. Also, there is an issue with missing data. Thus the first option is chosen here for consistency and simplicity. There might be a better solution to handle such circumstances and will be noted in the future work chapter.

Another issue regarding the time selection is the time format to be used. Since the time will be used to label the identity theft and to predict future trends, it is better to get the format of the time as specific as possible. However, news stories use various formats for the date. Some are more specific than others. It is hard to obtain accurate date information for each article. Thus a month and year format is chosen to represent the time the identity theft happens.

Based on the previous discussion, this research obtains the date from the story content and a month and year format are chosen for representing the time of the identity theft occurrence.

If a noun phrase is categorized as “Date” by the named entity recognizer, the phrase will be transformed to the standard MM-YYYY format and a further check is examined to eliminate the obviously invalid date such as May 1845. Then the date will be stored into the “Date” entry of the DAT file.

3.5.3 Finding the Location

The location of the identity theft is also important for the further analysis. Location will be used to analyze which states are at most risk for identity theft and the accumulative losses in those states. The location obtained by the named entity recognizer

is usually accurate and can be directly stored into the “Location” entry in the DAT file. Because only the state information is needed for further analysis, only the state name is stored. There are some similar problems to the time selection issues, such as multiple instances of location and missing location information. These problems are handled in the similar way as the time selection. Multiple locations will be weighted equally and the location entry for missing locations will be marked as not available. While a more accurate location could have been used, this requires the zip code instead of just the county and state name because multiple counties may use the same name within a state. Thus only the state name is chosen.

3.5.4 Risk Calculation

Risk is calculated for each identity attribute and based on the frequency of occurrence for a particular PII attribute in the news article. The predefined PII attribute list described in section 3.4 is used here. However, as mentioned earlier, this PII attribute list is not a complete one. Thus, a notion of potential attributes is introduced. 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. New PII attributes can be found by manually examining the potential attributes, or using some data mining techniques to classify the potential attributes.

For each predefined PII attribute, the occurrence is counted as the news story is being processed. Every noun phrase in the text is checked to determine if it matches a predefined PII attribute in the attribute list. If it is in the predefined list, the count for this attribute will increase by one. Otherwise, it will be stored in the potential attribute list. The frequency of occurrence for a PII attribute implies the probability this attribute is

exposed in the described identity theft. Thus, an attribute with higher frequency of occurrence in the news story will have a higher risk of exposure. This reasoning may not be true for all the news stories since in some cases, the low frequency of an attribute does not mean the risk of exposure is low. For example, in the analysis, the zip code may not have a high frequency of occurrence in the story. However since it is so widely used by people and easily obtained by the thief, it should have a high overall risk of exposure instead of a low risk. Consequently, it is important to explicitly state that the calculated risks are limited to the content of included news stories. The possibilities of quantifying the correlation between the word frequency and the risk of exposure are mentioned in the future work chapter.

3.5.5 Loss Calculation

Loss is calculated based on the ‘money’ name entity occurrences in the news article. The format of the loss obtained by the name entity recognizer could appear in several different ways. A format transformation is necessary to get the 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 will be transformed to the form of pure numbers.

First, the loss showing up in a single news story is added up and output as a sum loss. This sum loss is the loss for this particular story. In order to analyze the loss experienced by exposure of a particular PII attribute, some form of weighting is needed. Here, only the attributes matched in the predefined list are considered. An attribute with higher frequency of occurrence in the news story will have a higher weight. This is based on the reasoning that attribute frequency in an identity theft story indicates its importance

in the theft process. The formula below is used calculate the weighed loss for each attribute.

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

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

3.5.6 Timeline

The previous calculation for loss is only for a single dimension. In order to observe the trend of the change of loss, the correlation between the loss and the date are introduced. The idea is quite straightforward. 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 would be:

$$Loss_{a_i,d_i} = \frac{Loss_{sum} * f_{a_i}}{\sum_{i=0}^{n} f_{ai} * N_d}$$

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 for date d_i; f_{ai} is the frequency of occurrence for a particular PII attribute; N_d is the total number of dates in the story.

3.5.7 Theft Sequence Generation

How does identity theft happen? What steps are taken by the identity thieves? Generating the sequence of steps which thieves take and analyzing the correlation between different steps will help us to answer such questions.

3.5.7.1Step Representation

The first question follows. How should the steps in the identity theft scenario be represented? What is most important in the process? The actions the thieves take are represented in a graph to describe the sequence of steps in the thieves' business process and the resource/attribute involved at each step. There are two kinds of nodes in this graph, which are 'action' node and 'resource/attribute' node. Here the resource and attribute are treated in the same way due to the difficulty in distinguishing those two things in many circumstances.

This research addressed the questions: How should actions be represented? Should each unique verb be an action? Or should some sort of abstraction be used to represent the verbs? Since the sequence of the criminal's process steps is going to be compared among different identity theft stories, the unique verb representation would make it hard to identify the common behaviors the thefts may share. Thus the abstraction representation is used.

An action can be categorized into one of the seven categories: Record, Communicate, Decide, Act, Coordinate, Analyze and Collect. These seven categories of action are used to label each action. A bag of words that consists of the most frequently occurring actions in the identity theft news articles/stories are built and each action in it is organized into the corresponding category. Also a dictionary which consists all the tenses of those actions is generated to assist the mapping and organization of actions.This is done by using Simplenlg [34] which is a simple Java API (Application Programming Interface) for natural language generation. This approach uses the stemming which reduces the different forms of the same word to the same stem. For example, 'steals', 'stealing', 'stole', 'stolen' are regarded as the same as their stem, 'steal', and are all categorized as an "Act" action.

3.5.7.2 Generating steps

A story line can be generated by using the typed dependency parser from the Stanford's CoreNLP library to parse the input story sentence by sentence. For each sentence, the 'dobj (direct object of the Stanford NLP)' is identified. The direct object of a verb phrase is the noun phrase which is the object of the verb. For example, consider the sentence "He stole my credit card". After parsing, the dobj, (stole, credit card) is identified. 'stole' is the action and it would fall in the category of 'Act' and 'credit card' is the noun phrase as well as the resource/attribute. In order to obtain a good abstraction of the story, the noun and verb extracted by dobj are checked against a predefined dictionary mapping of resources to categories and represented in a sequence graph, which is also known as a process diagram [see Section 4.2.7]. The sequence graph will show the steps the identity theft takes from the beginning, with little information and resources, to the end, stealing the victim's property.

3.5.7.3 Visualizing steps

Next, the extracted criminal process steps are input to the Prefuse visualization toolkit [35], which is an open source package for creating rich interactive data visualizations. A process diagram is created to help visualize the sequence of steps as the noun and verb pair abstractions. SQUARE nodes in the graph represent the action (verb) and each such node has an associated TRIANGLE node representing the resource used (noun). The nodes are colored according to their categories.

Chapter 4: Results and Analysis

This chapter will describe the result for running the proposed algorithm on more than 3500 identity theft related news stories collected from various news feeds. Also, a more in-depth analysis on several different aspects about the results is illustrated. Last, but not least, the results and analysis are discussed regarding an understanding of the Identity theft process and prediction the identity threat in the future. It is worth mentioning that the statistics in this chapter are only based on the data obtained from news stories. They don't mean the actual identity theft statistics across the country.

4.1 INPUTS

The first step is to define a set of key words that are highly related to identity theft in order to collect the identity theft news stories. A list of words was chosen from several candidate phrases based on manually observation of the samples gathered from searching results of the identity theft news stories. Table 1 lists the news Rich Site Summary (RSS) URLS obtained by searching the selected words. Each Google news RSS contains 100 original links to the identity theft news stories, which are used to collect the stories on a daily basis. The New York Times News RSS also provides several stories each day. However, an obvious problem here is that the links from different RSSs could have duplicate ones, which needs to be eliminated from the links set. Thus a “Hashset” is used to keep track of all the previous collected news stories links. Due to the duplicate and invalid ones, the actual number of the stories collected each day is not as large as it seems to be. While about 300 stories are each day on average, only around 40 valid stories are obtained after eliminating the duplicate and invalid links. One thing 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 talking about how to protect you from identity theft could be gathered using this search. These types of articles could be eliminated by using a more restricted search.

The Identity Theft Resource Center [30] is another source for identity theft stories. Their breach report consists of data breaches that are gathered from a variety of media sources and/or lists from state governmental agencies. The ITRC report is updated daily.

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
NewYork Times News	http://topics.nytimes.com/top/reference/timestopics/subjects/i/identity_fraud/?rss=1

Table 1 News RSS URLs Based on Identity Theft Related Keywords

4.2 RESULTS

Next, the stories are processed by using natural language processing and text mining techniques described in chapter 3 to create an identity theft record was generated for each story. Although the stories provide a fairly large amount of information, most of

the stories are lacking some elements. Thus, those identity theft records are sparse and will affect the results of the analysis in accuracy. Then, several interesting aspects of the identity thefts are analyzed below to help us better understand the identity theft process and build the basis for the future identity theft threat prediction..

4.2.1 Impacted Target

This research defines 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 government agency. Identifying the impacted targets is beneficial to know since it is an indication of the kind of people and entities the thieves target.

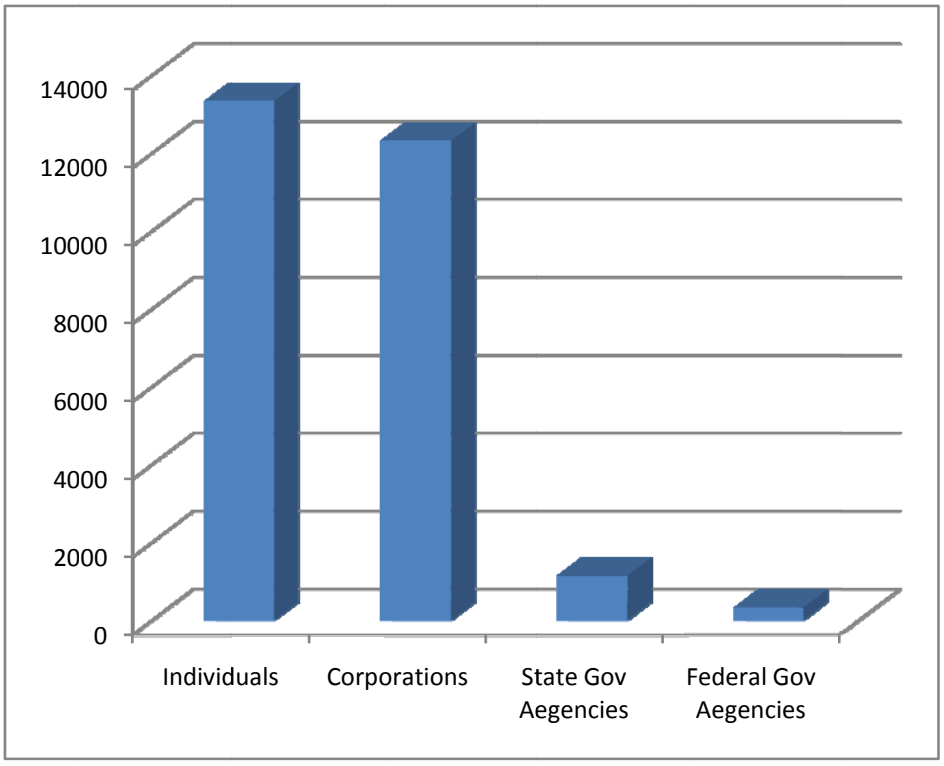

Figure 15 Impacted Target

Figure 15 presents the estimated numbers of the impacted target in the identity theft stories investigated. The news stories are processed using name entity recognition and the impacted targets are identified and categorized to the corresponding types. From Figure 15, we can see that individuals are most frequently targeted by identity thieves. The corporations are also targeted since a successful theft on a corporation could bring the thieves significant financial interests. If a corporation is breached and the criminal gets individual's data, both the individual and the corporation are considered as the impacted target here. The government agencies are not targeted as often as previous two probably because it is difficult to steal money from these government agencies.

4.2.2 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/resource is positively correlated with the frequency of it being used. Detecting commonly used resources could help to understand how fraudsters are gaining access to sensitive personal information. What are the most common attributes/resources being used to commit identity theft and fraud? fraud? Can access to these very common attributes/resources be limited or, at a minimum, make the providers or issuers of these attributes/resources aware that they are in fact being used in a malicious manner? Could this information be used to educate citizens about the importance of these attributes/resources and provide more protection for them? Often times, these attributes/resources are the pinnacle to the completion of a step of the whole identity theft process, and ultimately the theft. Therefore the analysis done in this area could significantly advance identity theft detection and prevention. The statistics generated are shown in Figure 16. The number in the graph indicates the

frequency of occurrence of the attribute/resource in the news stories. As indicated by the figure, social security number is the most used attribute. The SSN is important for almost every aspect of the lives of U.S. citizens. It can be used to apply for driver licenses, credit/debits cards and various kinds of applications. In some scenarios, the thieves even use the victim's SSN and combine it with a name and a birth certificate to forge a new identity which may appear as a brand new file on the credit bureau's record, which is extremely hard to track. Thus SSN has the highest risks among all the attributes. Credit/debit cards are also under high risk. Since almost everyone uses their credit/debit card every day, it has the high risk that the thieves could steal the card and use it directly to get the money.

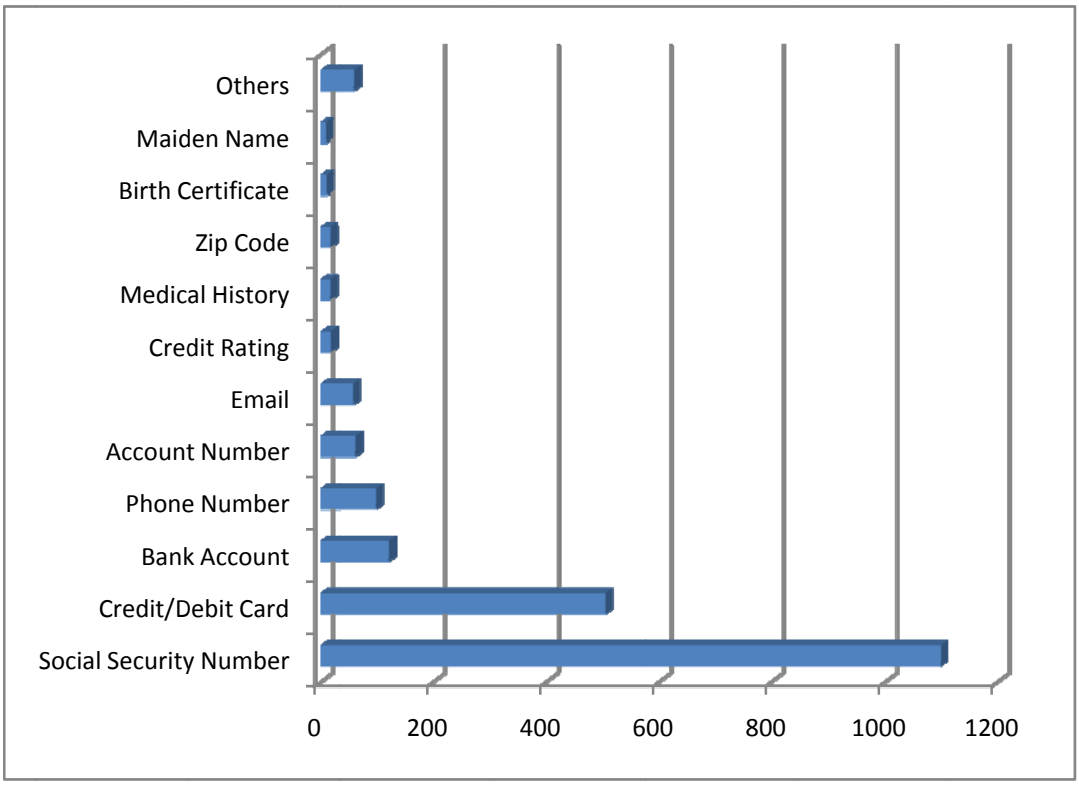

Figure 16 PII Attribute Risk Chart

4.2.3 Market Sector

Identity theft happens across a variety of areas and industries. The market sector here refers to the areas/industries where the identity theft occurred. Analyzing which market sector has the most identity thefts could help to understand the motivation of identity thieves and why they choose a given market sector as a target. Is it because this area is most fragile? Or is it because they can get the most financial outcome from this area? Can more recourse be invested to protect a person's PII related to this area? The occurring frequencies categorized by the market sectors are shown below.

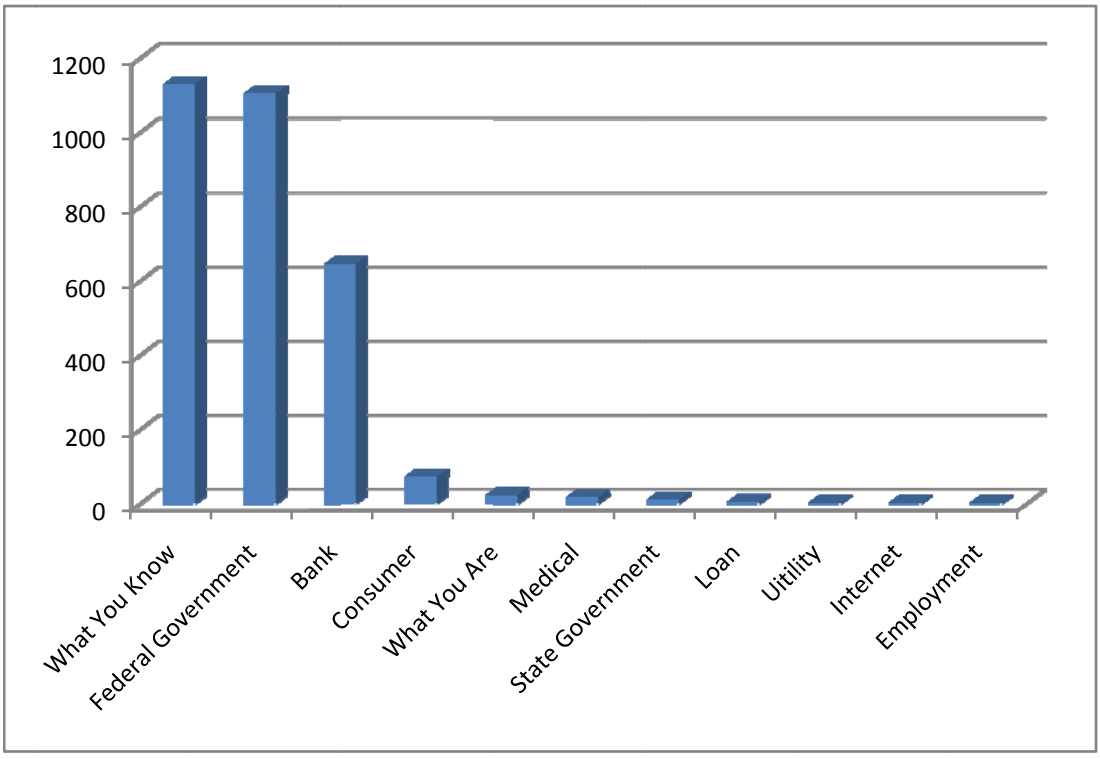

Figure 17 Market Sector Distributions

Figure 17 shows the market sector distributions in the identity theft stories investigated. As indicated by Figure 17, the top 4 market sectors consist of more than 90

percent of the total thefts investigated. Thus protecting those areas is very important to preventing the loss of people's financial interests.

In figure 17, "What you know" refers to the resource/attribute that an individual has knowledge about, such as your mother's maiden name, your home address, your bank password. "What you know" is a quite important category of PII and sometimes is only known by an individual. Thus if the thieves somehow gets to know something only known by an individual, such as one's bank account password, they can use such information to pretend to be that person and obtain benefits and resources.

The federal government category in Figure 16 refers to the attributes that are issued by the federal government, such as social security number, visa, etc. These attributes are also valuable since the thieves can use these to forge a new identity. The next two market sectors that have highest losses are bank and consumers services. Both are highly related to the financial interests of the victim. Bank and consumer services refer to all kinds of attributes that are related to the banking and consumer industry. These two areas have the third and fourth highest occurring frequencies probably because both market sectors are directly related to financial resources.

4.2.4 Location Analysis

The location where the identity theft and fraud happens is also important to explore. Figure 18 below shows the identity theft location-wise distributions in the identity theft stories investigated. The figure indicates that the two states with highest frequencies are California and Texas, which makes sense because those two states have the highest population. They have the most people and thus have a higher frequency of identity theft. Frequency per capita would also be interesting to invetigate. However, the

data obtained in this approach does not rule out bias in terms of location, it is not very useful and accurate to calculate such information. Attempting to collect per capita identity theft should be pursued in future work.

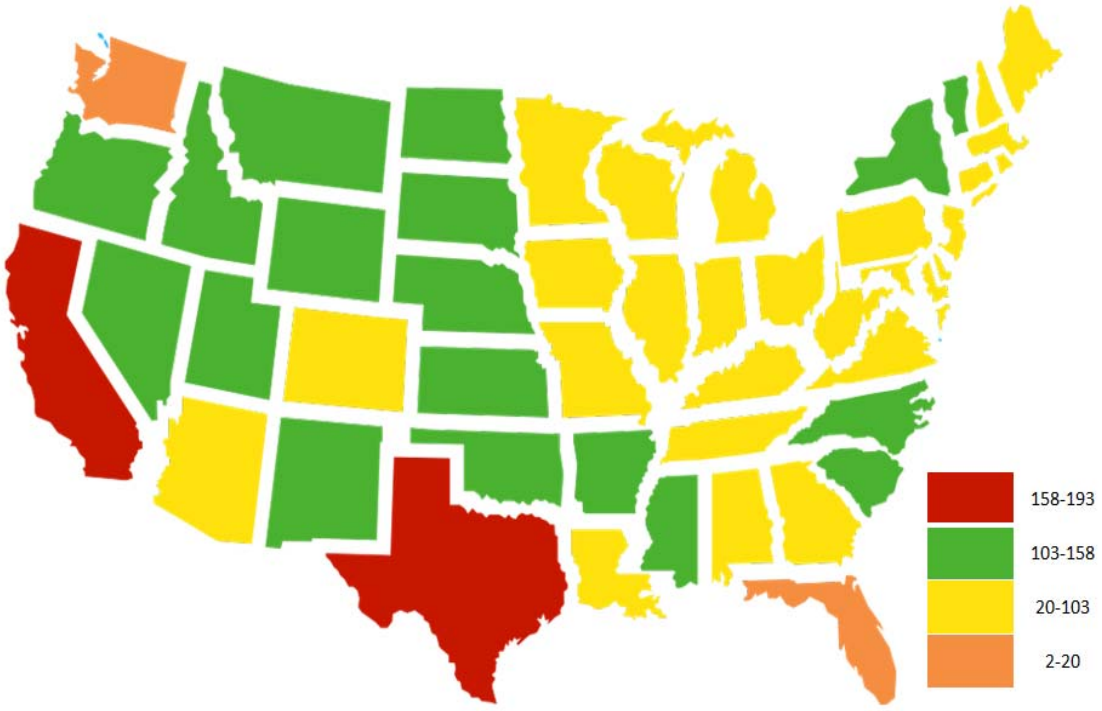

Figure 18 Identity Theft Map

4.2.5 Financial Impact Analysis

The financial impact of identity thefts is of significant concern. Although the identity thieves may use the victim's identity to commit a serious crime such as launching a terrorist attack, most thieves pursuing the financial interest behind the identity. Which PII attributes, if compromised, can cost the victim the most financial loss? Is enough attention put on such attributes? Should the investment on protecting these attributes be

increased? Analyzing such information may help to improve a person's awareness to protect higher valued attributes.

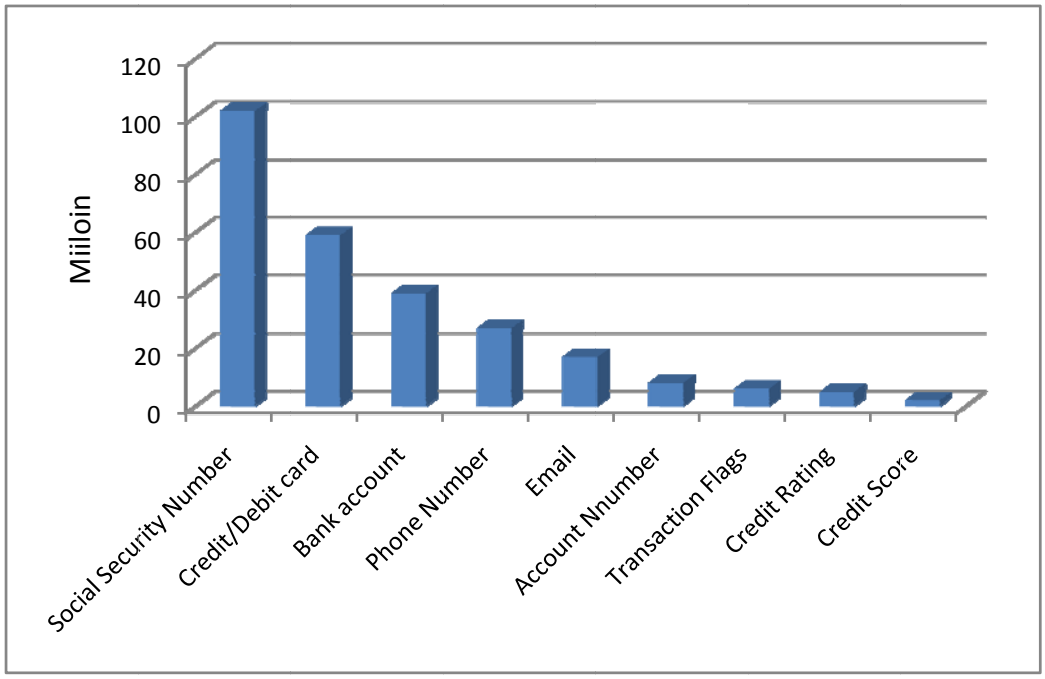

Figure 19 Financial Impact per Attribute

Figure 19 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 basically match the figure for the risk of all the attributes. The social security number caused the most loss among all attributes. As discussed earlier, this is probably because SSN is important for almost every aspect of our lives. The credit/debit card and the bank account are still the second and third. Moreover, the phone number and email here rank fourth and fifth among the most costly attributes, which matches their high risks in Figure 16. At a first glance, it seems that the phone number and email should not cost so much loss since they don't have a financial value themselves. However, considering they are

widely being used for authentication for all kinds of financial accounts. The identity thieves could use these attributes to pass the authentication and reset the password or security questions, which can indirectly get access to those accounts and stealing money.

4.2.6 Timeline Analysis

The previous analyses are solely based on the frequency or total loss related to a particular attribute. What if adding the time as a factor to conduct those analyses? The timing information will help us to better understand the trend of the attribute values changes. For example, smart phones are becoming more and more popular and people start to use these phones to store valuable information and pay for bills. The phone could even be used to identify a person now since one could access their bank account by using their phone. Many websites also use text messaging confirm one's identity. Thus the phone's value has increased as time goes.

Figure 20 presents the monthly loss related to the 5 attributes that cause the most losses in the identity theft stories investigated. The figure shows very few losses before the year 2011. This is because the news stories are mainly collected from the recent year's news. Only a few of stories describe identity theft story before 2011; therefore, the trend before 2011 seems random. There are also several peeks showing up at the end of 2013 in the figure. One possible reason -- the Target data breach happened in mid December 2013. The criminals forced into Target's system and gained access to guest credit/debit card information. A lot of news stories reported the identity theft related to the Target data breach. Therefore, the loss around that time has a peak value. However, there is no hard evidence indicating that the peak value is caused by the Target data breach. This is just a speculation. Further observation about the data and original news

story may be helpful to understand the reason of the increase. It is worth noting that many points are missing along the timeline, even in recent months. This could be due to that the timing information for those stories is missing, which is either because those stories don't have a valid date that can be extracted in a good format or they just simply don't have a date.

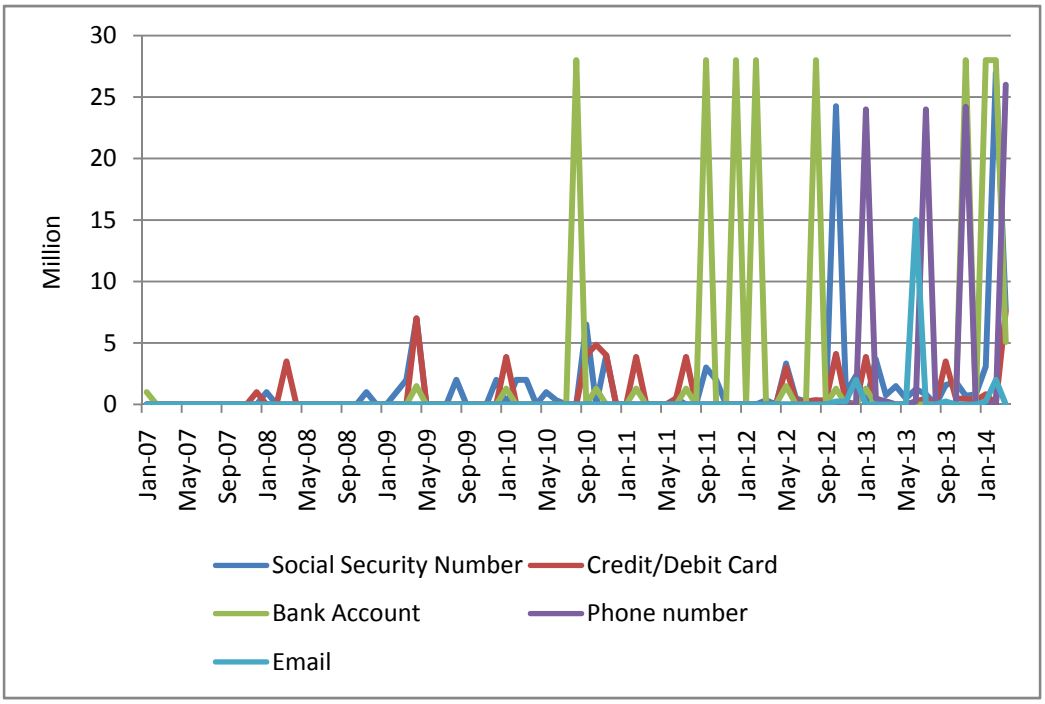

Figure 20 Loss per PII Attribute per Month (Top 5)

Figure 20 describes the Loss per PII attribute in a monthly view. Figure 21 below shows an accumulative view of the same data. The figure indicates that the loss associated with these attributes is increasing faster recently. Does this mean that the loss caused by the identity theft is increasing? It partially props this perspective. However, since the data set being used is biased towards recent thefts. This conclusion cannot be made without more in-depth examination. Another thing can be observed from this figure is that the trend for each attribute coincides with each other roughly. This implies that

these attributes are all used very often during this time period. However, if more attributes are included in this figure, such implication may not be true. Some attributes are not showing here is because the total loss caused by these attributes is not high enough to be in the top five costs. The figure also indicates that the loss caused by a phone number increased very fast in recent months, more than other attributes, which can help us to predict phone number future value to criminals and should forecast increased protection of phone to combat future identity thefts.

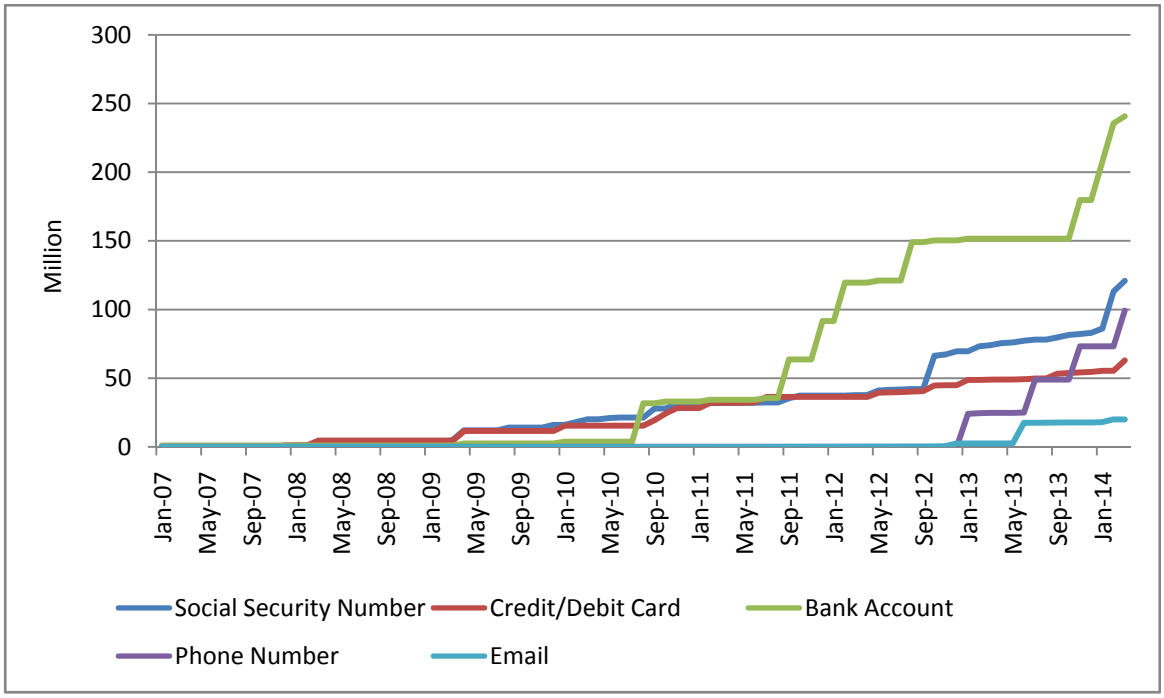

Figure 21 Accumulative Loss per PII attribute (Top 5)

From the previous two figures, one can clearly see the sparse characteristics of the current data set. More stories are needed for a more thorough analysis. The data used for this thesis is obtained from gathering the news stories for forty five days on a daily basis as well as from the breach report generated by the Identity Theft Resource Center [33].

More data will increase confidence in the results. Also, these timeline figures could be used to predict risk and cost trends related to a particular attribute, which can help us protect these attributes and get a step ahead of the thieves.

4.2.7 Process Diagram Example Analysis

How is identity theft implemented? What steps are taken by the identity thieves? Generating the sequence of steps, i.e. the process diagram, which the thieves take and analyzing the correlation between different steps will help us to answer such questions. Figure 22 shows an example of the process diagram for the Home Equity Fraud scenario.

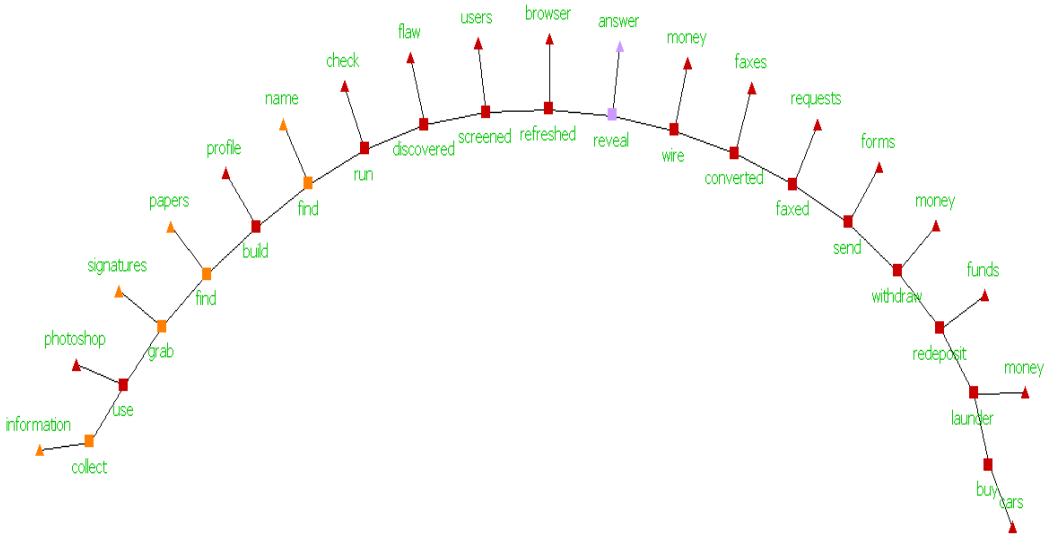

Figure 22 Home Equity Fraud Process Diagram

The input article is an identity theft story from the ITAP database called “Home Equity Fraud”. The figure shows a simple outline of part of the story and from it one can

get a rough idea about the whole process of the theft. Firstly, the fraudster collected mortgage information on wealthy couples and searched for lease and loan documents in public databases. Then, he used tools to grab signatures from the loan documents and built a profile of the victim. Next, he discovered a flaw in the Experian portion of the site which is refreshing the browser enough times would reveal the true answer of the security questions to access reports. Then he wired money out of the country and had someone withdraw the money and redeposit the funds into other account. He would further launder the money by depositing it. However, the sequence described by the process diagram doesn't provide all the information about the identity theft story. Also it neglects some details compared to the whole story narrated in the second chapter. There are two major reasons. First, the article is a story. It is not an investigative report about the identity theft. So the article doesn't have enough details of how the identity thieves committed the crime and which tools they use. Secondly, the proposed algorithm doesn't work well if the sentence becomes too complicated and uses a phrasal verbs instead just a single verb. In addition, finding out and extracting the tools/techniques the thieves used to steal a person's identity are also a useful but challenging task. But in an article, it is hard extract such information given the ways the author used to describe such information -- using a verb object phrase or a prepositional phrase, or even worse, the author doesn't describe this information at all.

What other information needs to be extracted from the news story and from the process diagram? First, the research effort is seeking to find the most common data inputs required for any given step within a scenario to occur. Establishing the most common data inputs required to complete a capability is critical to understanding and potentially thwarting an identity attack. Without this initial piece of data, many attacks would not

take place. What information did the thieves already have on hand? Did the thieves have access to your email account? Were they able to view your place of birth or birth date on Facebook? By understanding what the thieves possess in the beginning, it is possible to change the way consumers feel about certain elements of data which they, until now, thought to be secure. Perhaps a mother's maiden name is no longer a secure method of identifying someone since it can be easily discovered by viewing family trees on Facebook or Ancestry.com. With the proper data showing patterns of repeated identity theft resulting from a certain data element, the research group can hopefully make the general public aware that perhaps a different data element should be used to identify themselves. Most people do not currently feel the need to safeguard their phone number or email address. Based on this research, better education can be provided regarding the protection and security of personal identifiable information.

Second, this research effort wants to understand exactly what the fraudster is after in each given scenario. Showing the most common data output resulting from certain steps in the identity thieves' business process helps us create a full picture of the entire business process. Often, the data output in one capability is a necessary input to the next step. What are the most common data outputs? It is important to ascertain exactly what the fraudster hoped to accomplish at each step. This will provide a better understanding of exactly how the criminal is getting their hands on certain pieces of information. How exactly did the Home Equity fraudster end up with the victim's Mother's Maiden Name? Name? He was able to Google information and then easily use it as a data input on Ancestry.com. Perhaps it is time to consider different data inputs as a method of identifying someone, especially on sites such as Ancestry.com, where the entire purpose of the site is information discovery. This shift in our thinking may be especially

important if the data being acquired could be potentially detrimental to someone's personal financial, physical or emotional security.

Furthermore, the research effort aims to detect any repetitive groupings that exist between steps in the scenarios. This could be done by finding patterns from various scenarios. Understanding if there are groups of capabilities that often work together, may help us visualize patterns emerging once a string of events occur. This is important because it could serve as a prediction tool for future identity thefts. Are there two or three steps that typically go together across multiple scenarios? Could these steps, when completed and detected in conjunction, throw a red flag that identity theft may be occurring? It is certainly possible. Because the ITAP make connections between steps, the tool can be used to explore what steps commonly work together in the process of committing identity theft and fraud. Does step A typically follow step B? And how can people use knowledge of this to prevent the next step C?

Chapter 5: Conclusion

The main motivation of this thesis is gather, model and study the data necessary to analyze and predict behaviors of identity thieves and fraudsters. This research collected news articles from the Internet. Using the text mining techniques, criminal behaviors were analyzed and formulated as a basis to predict future trends of identity theft and fraud. The whole process was automated after the initial setup stage for the predefining PII attributes and categorizing the behavioral actions into seven categories: Record, Communicate, Decide, Act, Coordinate, Analyze and Collect. The system is also designed in a pipelined fashion where each step can be done separately and integrated together to build the system. The proposed algorithm is to identify new attributes for enriching the attribute list. This step currently involves human examination and selection before new attributes are added to the attribute list.

This research employed an approach for mining the news stories similar to the existing techniques used for mining general text. The first step is to obtain the news stories from the Internet. Around 3500 identity theft news stories were gathered. Story text is preprocessed and irrelevant and unnecessary information is eliminated. After that, the named entities are extracted by using the named entity recognizer. These named entities are then 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 have experienced the identity theft, the risk for losing a particular PII attribute, the frequency of identity theft's occurrence in different market sector, the location where the identity theft happens, the potential financial impact caused by a comprised PII attribute, the changes of such impact along with the time. Analysis of these results should help researchers to better understand

identity threat behaviors, offer people early warning signs and thwart future identity theft crimes.

Additionally, the sequence generation is done by parsing each sentence and finding the typed dependency between different components within a sentence. The dependencies are used to generate the process diagram, which can be used to better understand the identity theft process.

Chapter 6: 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 thesis, 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 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.

Secondly, 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, or just be assigned with different weights?

Third, this thesis 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 just don’t occur often in the identity

theft news stories. For example, the 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.

Fourth, how include the new attributes to enrich the predefined attribute list is a hard task. This thesis generates a potential attribute list for each news story. However, it still needs to be manually selected from this list to get the new attribute. One automated approach to build a model and use the current attribute and news stories to train this new model. Then use the new stories as the input and generate an attribute list. Future research should also consider the maximum words related to a single attribute as well as how the system deals with multi-word phrase. To identify whether a multi-word phrase is an attribute is much harder. An N-gram model may help solve this problem.

Fifth, instead of using the news story, one can use a more informational text as the input to the system, such as an identity theft victim stories or law enforcement reports. These stories may have more accurate and complete information and offer a more structured, sequenced story.

Last but not least, where else could the data generated by the system be utilized? One obvious application is that to use this data to feed the ITAP system. ITAP system could improve its model and predict the trend for different aspects related to the identity theft. For example, it can be used to predict whether a resource/attribute is at increased risk and whether the thieves will use it more often.

References

[1] Graeme R. Newman and Megan M. McNally. *Identity Theft Literature Review*. National Institute of Justice, 2005

[2] AWARE Software, Inc. *AWAREnes User Guide*. AWARE Software, Inc., Austin, Texas, 2014.

[3] Harmon, Paul. *Business Process Change: A Guide for Business Managers and BPM and Six Sigma Professionals.* 2nd ed. Amsterdam: Elsevier, 2007.

[4] Witten, Ian H., Katherine J. Don, Michael Dewsnip, and Valentin Tablan. "Text mining in a digital library." *International Journal on Digital Libraries* 4, no. 1 (2004): 56-59.

[5] Aase, Kim-Georg. *Text Mining of News Articles for Stock Price Predictions*. Norwegian University of Science and Technology, 2011.

[6] Wang, Yi, and Xiaojing Wang. "A new approach to feature selection in text classification ." Machine Learning and Cybernetics 6 (2005): 3814 - 3819.

[7] Al, Ahmad, and Qasem Abu. "IRS for Computer Character Sequences Filtration: a new software tool and algorithm to support the IRS at tokenization process." International Journal of Advanced Computer Science and Applications 4, no. 2 (2013): 81-82.

[8] Feldman, Ronen, and James Sanger. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge: Cambridge University Press, 2007.

[9] Baker, L. D., and A. K. McCallum. "Distributional clustering of words for text classification." *In Proceedings of SIGIR-98, 21st ACM International Conference on Researchand Development in Information Retrieval* (1998): 96-103.

[10] Porter, M.f.. "An algorithm for suffix stripping." *Program: Electronic Library and Information Systems* 40, no. 3 (2006): 211-218.

[11] Wang, Xuechuan, and Kuldip K. Paliwal. "Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition." *Pattern Recognition* 36, no. 10 (2003): 2429-2439.

[12] Liu, Huan, and Hiroshi Motoda. *Feature selection for knowledge discovery and data mining*. Boston: Kluwer Academic Publishers, 1998.

[13] Bogdanova, Dasha. "Extraction of High-Level Semantically Rich Features from Natural Language Text." *ADBIS* 1 (2011): 262-271.

[14] Mooney, Raymond. "Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning." *In Proceedings of the Conference on Empirical Methods in Natural Language Processing* (1996): 82-91.

[15] Serrano, J.I., and L Araujo. "Evolutionary algorithm for noun phrase detection in natural language processing." *Congress on evolutionary computation (CEC)* (2005): 640–647.

[16] Lester, Mark, and Larry Beason.*The McGraw-Hill handbook of English grammar and usage*. New York: McGraw-Hill, 2005.

[17] Ghahramani, Zoubin, Thomas L. Griffiths, and Peter Sollich. "Bayesian nonparametric latent feature models." *Proc. Valencia / ISBA 8th World Meeting on Bayesian Statistics* (2006)

[18] Deerwester, S., G.W. Furnas, and T.K. Landauer. "Indexing by latent semantic analysis." *Journal of the American Society for Info, Science*41 (1990): 391-407.

[19] Strzalkowski, Tomek. "Document representation in natural language text retrieval." *In Proceedings of the Human Language Technology (HLT) Conference* 0 (1994): 364--369.

[20] Salton, G., A. Wong, and C. S. Yang. "A vector space model for automatic indexing."*Communications of the ACM* 18, no. 11 (1975): 613-620.

[21] Salton, Gerard, and Michael J. McGill. *Introduction to modern information retrieval*. New York: McGraw-Hill, 1983.

[22] Ramos, Juan. "Using tf-idf to determine word relevance in document queries." *In Proceedings of the 1st Instructional Conference on Machine Learning* (2003).

[23] Sang, Erik F. Tjong Kim and Fien Meulder. "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition." *Proceedings of CoNLL-2003* 1 (2003): 142–147.

[24] Lafferty, J., A. McCallum, and F. Pereira. "Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data." *In Proceedings of the 18th ICML* (2001): 282–289.

[25] Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. "Incorporating non-local information into information extraction systems by gibbs sampling." *In Proceedings of ACL* (2005): 363–370.

[26] Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. "Building a large annotated corpus of English: The Penn treebank." *Computational Linguistics* 19, no. 2 (1993): 313–330.

[27] Toutanova, Kristina, Dan Klein, and Christopher Manning. "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network." *In Proceedings of HLT-NAACL* (2003): 252-259.

[28] Marneffe, Marie-Catherine de, Bill MacCartney, and Christopher D. Manning. "Generating Typed Dependency Parses from Phrase Structure Parses." *LREC* (2006): 1.

[29] Golden, Ryan. *NewsFerret: Supporting Identity Risk Identification and Analysis Through Text Mining of News Stories*. Master Report: The University of Texas at Austin, 2013.

[30] Identity Theft Resource Center. (2014, March). Data Breaches.

[Online]. http://www.idtheftcenter.org/id-theft/data-breaches.html

[31] boilerpipe. (2014, March) boilerpipe: Boilerplate Removal and Fulltext Extraction from HTML pages. [Online]. https://code.google.com/p/boilerpipe/

[32] Apache PDFBox. (2014, March) Apache PDFBox.
[Online]. http://pdfbox.apache.org/

[33] The Stanford NLP (Natural Language Processing) Group. (2014, March).Stanford CoreNLP.[Online]. http://nlp.stanford.edu/software/corenlp.shtml

[34] SimpleNLG. (2014, March). SimpleNLG.
[Online]. https://code.google.com/p/simplenlg/

[35] Prefuse. (2014, March). Prefuse. [Online]. http://prefuse.org/