# *NewsFerret*: Supporting Identity Risk Identification and Analysis Through News Story Text Mining

Ryan Golden
Center for Identity
Department of Electrical & Computer Engineering
The University of Texas at Austin
Austin, U.S.A.

Suzanne Barber
Center for Identity,
Department of Electrical & Computer Engineering
The University of Texas at Austin
Austin, U.S.A.

*Abstract*— **Individuals, organizations, and devices are now interconnected to an unprecedented degree, forcing identity risk analysts to redefine "identity" in such contexts and explore new techniques for analyzing expanding threat contexts. Major hurdles to modeling in this field include a lack of publicly available data due to privacy and safety concerns, as well as the unstructured nature of incident reports. Thus, this report uses news story text mining to develop a new system for strengthening identity risk models. The *NewsFerret* system collects and analyzes stories about identity theft, establishes semantic relatedness measures between identity concept pairs, and supports analysis of those measures with reports, visualizations, and relevant news stories. Risk analysts can utilize the resulting analytical models to define and validate identity risk models.**

***Keywords-identity; risk analysis; risk modeling***

## I. INTRODUCTION

Identity theft, fraud, and abuse incidents affect large numbers of citizens, small businesses, corporations, and government agencies around the world. Between 2004 and 2011, 11.6 million adults were victimized in the U.S. alone, resulting in $18 billion in financial losses [1]. And while many traditional identity threats still persist (e.g., mail theft), the modern identity threat space has evolved rapidly in recent years as individuals, organizations, and devices are increasingly connected around the world.

In 2012, the number of internet-connected devices around the world grew to above 9 billion [2] and successive waves of innovative technologies (e.g., e-commerce, smart phones, tablets, social media, big data, cloud services) are exposing new inroads for malicious actors. New identity attributes, such as check-in location [3], now exist and lay exposed. Furthermore, well-established identity attributes such as credit card number and social security number face new threats as they are digitized and replicated in global big data systems [2].

To help understand, assess, and control these identity threats, the Center for Identity (CID) at the University of Texas at Austin is developing systems to assist with identity risk modeling. These systems provide a framework for studying the structure of an identity, its areas of vulnerability, and the estimated costs to personal, financial, or national security when elements of its structure are compromised. CID is also creating an analytical repository of known identity threats and counter measures, structured in a suitable form for analysis [4]. Acquiring and processing enough data on identity threats to validate models of identity risk is a crucial challenge in developing these systems. Thus, this report illustrates and evaluates a practical solution to this problem—a system called *NewsFerret*—that collects, models, and analyzes large numbers of news stories on identity threats.

## II. IDENTITY RISK MODELING AND ANALYSIS

Organizations employ *identity risk modeling* techniques to combat and manage identity theft, fraud, and abuse threats. Possible scenarios where identity risk modeling is important include: a company or government agency must choose how to focus resources in response to an identity theft or data breach incident; law enforcement conducting forensics on an identity theft must identify potential suspects or attack vectors; and a software or network architect must design a system that will store, authenticate, and use identifying information.

In each of these scenarios, organizations seek to protect *identity attributes*, or the intrinsic elements of an identity or set of identities within a particular domain that have value. For example, the identity attributes of individuals who interact with an e-commerce web site may include user login, e-mail address, credit card number, mailing address, and password. An *identity ecosystem* is a collection of these identity attributes plus all of the relationships between them. For example, a password that grants access to a bank account or a social security number that authorizes a tax return.

The actors, actions, and resources that serve to steal, defraud, or abuse identity ecosystem elements compose a *threat context*. For example, a hacker may use a fake credit card number generator to gain access to an Amazon account, or an identity thief may steal mail from mailboxes to forge a job application. The combination of all possible threats makes up the threat context.

The abstract region where an identity ecosystem and a threat context overlap is referred to as the identity risk of a domain. Organizations dealing with identities naturally work towards reducing identity risk as a means to reducing cost and liability. Where risk cannot be eliminated, organizations—

more specifically, analysts within organizations—seek to understand and control it through a set of steps using an identity risk model.

The *identity risk analyst* is interested not only in privacy risks, but also in risks to systems, resources, and reputation. Responsibilities of the risk analyst include:

- *Identify risk.* The analyst identifies which elements of an identity are at risk and should be part of the risk model. For example, should "Facebook account" be part of the risk model, or is that attribute not relevant to any known threat?
- *Analyze risk.* The analyst seeks to understand the mechanisms by which the identity elements are put at risk, and seeks to quantify the likelihood that an incident should occur. To support findings, the analyst validates the risk model against existing incident data.
- *Assess risk.* The analyst values stakeholder impact—e.g., cost or damages—of the risk, should an incident occur. For example, if a user's bank account is compromised, the cost is very high to both the bank and the customer.
- *Manage risk.* The analyst—typically in collaboration with a larger team—implements controls and countermeasures to mitigate, prevent, or counteract the risk. The analyst communicates the risk model to stakeholders, and keeps the risk model up-to-date.

To perform the above responsibilities effectively, the risk analyst must have a valid risk model. To create a valid risk model, the analyst may seek to answer the following questions:

- Given a set of compromised identity attributes within an ecosystem, what other attributes may be at risk, and to what degree?
- Given a set of compromised identity attributes within an ecosystem, what are the likely attack vectors (i.e., actors, actions, resources) that compromised the attribute? How likely are these vectors?
- Given a threat context, are the elements of my identity model valid? Are any elements missing in my model? For example, do threats exist to attributes I have not considered in my identity model?
- Given an ecosystem, are the elements of my threat model valid? Are any elements missing in my model? For example, do threats exist that I haven't considered in my threat model?

Thus, identity risk modeling involves identifying, analyzing, assessing, and managing risk. Valid risk models should represent—as faithfully as possible—the underlying, real-world, semantic structure of the identities and threats being modeled. The effort to capture this real-world structure through a model can be visualized by Fig. 1.

To validate risk models, analytical tools must be backed by a sufficient amount of data to support an analyst's conclusions. One possible source for this data is a repository of structured incident records. However, such a repository requires data entry by a domain expert familiar with the structured incident format. The identity threat scenarios modeled in this way for the Center for Identity's incident repository—the Identity Threat Assessment and Prediction (ITAP) system) [5]—are not yet sufficient in number. At the same time, there exists somewhat of an inverse problem. The global news media reports hundreds of identity incidents each month. Yet manual data entry of all of these reports into ITAP is not feasible due to their unstructured and voluminous nature. *NewsFerret* enables meaningful analysis given this dual problem: (1) too little structured threat data (incident repositories); and (2) too much unstructured threat data (news stories).

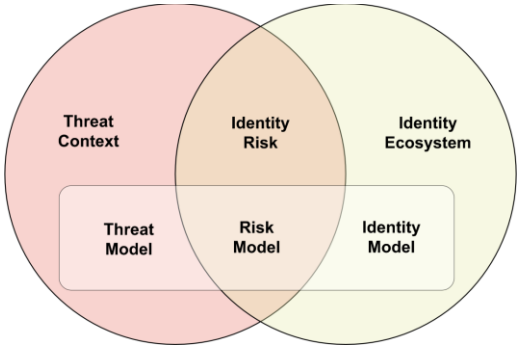


Figure 1. Modeling identity risk.

## III. Related Work

Before examining *NewsFerret*'s capacity to address the challenges of modeling real-world scenarios, a review of related work that informed the design of this system is presented. First, identity, privacy, and security are clearly related, yet distinct, fields. Privacy risk models often focus on the perspective of individuals rather than devices or organizations. Also, in privacy risk modeling, risk is related to an individual's social obligations or potential embarrassment rather than systems or resources. Additionally, adversaries in privacy risk scenarios often already know an individual's identity (e.g., parent, employer, friend, or acquaintance) [6]. Identity risk study casts a wider net to include devices, organizations, and abstract entities such as online personas, avatars, or profiles used in different systems.

Another related field is security threat modeling [7]. Security threat modeling is an important component of identity risk modeling, but not alone sufficient to model identity risk. Historically, security study has focused on "mechanisms and techniques that control who may use or modify the computer or the information stored in it" [8]. Modeling identity risk may mean defining or enforcing such security threat controls. However, identity threat modeling contrasts with security risk modeling in that an identity often comprises information across multiple computer systems and networks. Indeed, attributes of a single identity are often dispersed and replicated around the world, in systems both human and machine that are controlled by organizations with different interests.

Traditional techniques of modeling security threats must be updated to handle a more distributed, inter-connected, and dynamic definition of a protected resource. Security threat modeling is typically performed from the system designer's perspective to prevent an attacker or intruder from accessing protected data [7]. The identity risk modeler will also consider this, but may additionally be concerned with how to prioritize incident response after an incident has occurred.

Finally, The Center for Identity's ongoing project to develop Identity Ecosystem Maps (IEMs) [4] heavily informed this work, especially choosing to model identity ecosystems using attributes and relations. However, IEMs offer a framework for analysis, but not prescriptions for mechanisms to populate specific attributes and relation values. *NewsFerret* complements IEMs as one such data input mechanism. Specifically, *NewsFerret* realizes a text mining system [9] design that measures semantic relatedness between sets (e.g., pairs) of concepts present in the identity risk space. At a high level, the system collects news stories from the web based on a set of keyword-based topics (example keywords are "identity theft" and "identity fraud"), analyzes the news stories using text mining techniques, and produces reports and analytical models suitable for identity risk identification, analysis, visualization and export to other analytical models.

## IV. Hypothesis & Example Scenario

The hypothesis is that keyword topics found in news stories (e.g., "identity theft") represent some subset of the semantic structure of identity risk. By modeling this semantic structure and comparing it against a risk model, the analyst can understand where her model may be valid or invalid, or may be missing elements. To explore this hypothesis, the following hypothetical scenario is used throughout this report to illustrate how *NewsFerret* can help validate risk analysis and modeling:

Amazon.com users report an unusual number of recent credit card thefts. The attack vector is unknown, although internal security teams theorize the hackers are exploiting a widely reported technique [10]. First, the hacker calls customer support to add a fake credit card number, then calls again to add a new account e-mail address. This technique compromises the user account and leads the hacker to a user's credit card information. The strategy requires three user identity attributes: name, billing address, and e-mail address. The required hacker resources include a fake credit card number generator and the customer support phone line. Based on this theory, the internal security team has created a risk model (see Fig. 2). However, no one is sure of the model's accuracy or the value the risk posed by individual elements in the model; aside from the number of reported credit card thefts, there is no data upon which to base any conclusions. Consequently, as a security consultant specializing in identity risk modeling, you have been hired to lead a team of risk analysts to validate the client's identity risk model and answer the following questions:

- What are the likely attack vectors (i.e., elements from which the attack originated)?
- Are any elements missing in the risk model?
- What elements pose the greatest risk (i.e., where should the client expend energy and resources to mitigate risk)?

Your team now has all the pieces of information necessary to use *NewsFerret*: a description of the threat context (credit card information is being stolen), a description of the identity ecosystem (user name, credit card, e-mail, etc.), and a starting model. Subsequent sections of this paper revisit this example to illustrate how this information can be fed into *NewsFerret* to help validate this model, answer the above questions, and explain important design decisions.

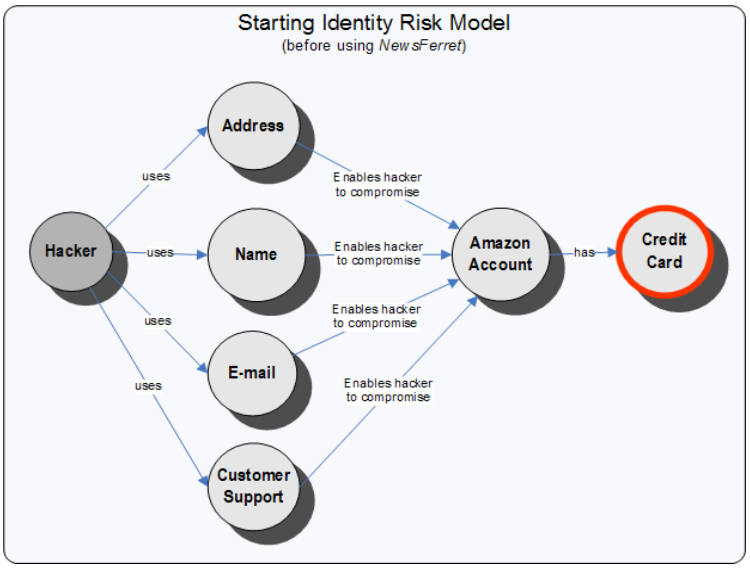


Figure 2. Modeling identity risk.

## V. DESIGN

The first major design decision, therefore, is utilizing news stories as stand-ins for the real-world, semantic space of the identity risk. To do so, the system must meet certain functional and non-functional requirements.

### A. Functional Requirements

The following step-by-step process illustrates the primary functional scenario the system should support to satisfy objectives, stakeholder responsibilities, and aspects of the Amazon example scenario.

*1)* The risk analyst declares a set of keywords matching the threat context of interest for modeling risk. In the example scenario, based on the threats under investigation, the analyst declares keywords "identity theft" and "credit card theft."

*2)* The analyst defines a starter list of identity attributes she is interested in understanding and analyzing for this threat context. In considering the list of attributes, the analyst may consider the attribute types shown in Fig. 1, based on discussions of multi-factor authentication [11]. In the example scenario, based on the starting risk model (see Fig. 2), the analyst defines the starter list to include "name," "address," "e-mail," "Amazon account," and "credit card."

*3)* The system gathers and stores information describing the given threat context. Typically, this information may be in the form of a description or report of a particular incident or a threat scenario occurrence. In the example scenario, the system will gather and store news stories matching the keywords "identity theft" or "credit card theft."

*4)* The system builds analytical objects that model identity risk within the threat context. For the example scenario, analytical objects and reports are described in the Analysis section of this report.

*5)* The analyst uses the analytical objects and services provided by the system to validate her existing identity risk model. In the example scenario, the analyst adds risk model elements based on new information in the analytical objects (a revised risk model is described in the Analysis section).

TABLE I. IDENTITY ATTRIBUTE TYPES

| ***Attribute Type*** | ***Examples*** |
|---|---|
| Things you are | Eye color, DNA, maiden name, IP address, federal tax ID, location |
| Things you have | Money, driver's license, credit card, bank account, Medicare card, brand, reputation |
| Things you know | Password, last year's tax return amount |

### B. Non-functional Requirements

The system must also satisfy non-functional requirements dealing with privacy, scalability, and timeliness. The inherent sensitivity of incident data is an important factor in identity studies and other security-related risks [6]. For example, in the U.S., incidents reported to most law enforcement agencies are only available to other law enforcement agencies [12], since incident details often reveal personally identifiable information (PII) that could be used by future attackers. Thus, the system must gather and store threat scenarios without revealing PII. Additionally, since access to this type of data is limited and difficult to gather, the system should ideally use a public data source.

The system should also be scalable. As previously discussed, manual entry of incident data into a structured format is both time-consuming and error-prone. To obtain the necessary volume of data for accurately portraying a threat context, the system should minimize manual data entry. Furthermore, the system should be able to store and process large amounts of textual data to support flexible system expansion.

Finally, the system should acquire up-to-date threat data. Like other security-related risks, identity security risks often show trends over time, and today's predominant threat scenarios may not be relevant tomorrow. It is important, therefore, that the risk analyst model risk based on incident data that is as timely as possible. Requiring manual entry of incident data inhibits this need; again, the system should avoid requiring manual data entry whenever possible.

### C. Text Mining System

The *NewsFerret* realizes a text mining system design by analyzing news stories with text mining techniques including latent semantic analysis [9]. The latent semantic analysis (LSA) technique [13] [14] is a form of dimensionality reduction that exhibits an interesting property whereby an approximation of the observed data is more informational than the raw observed data. This approximation represents the "semantic space" of concepts in a set of documents [15]. LSA and related techniques can be used for purposes such as similarity comparisons in the semantic space and information retrieval; *NewsFerret* uses techniques for both purposes. Similar techniques are sometimes referred to as "concept linkage" and used in other domains such as systems biology [9] [16].

### D. Operational Reference Model

This design uses an Operational Reference Model diagram [17]. The elements included in this diagram (see Fig. 3) are presented in process order:

*1) News Published:* A media outlet publishes one or more news stories on a topic relevant to the domain the analyst wishes to study. For example, a regional newspaper publishes an article, "Rome woman charged with identity theft," [18] which is subsequently aggregated by a news feed publisher under the keyword topic of "identity theft."

*2) Collect:* The system retrieves the news article(s), based on information from one or more news feeds; extracts the core content of the web page; and stores the news article content along with relevant metadata including author, publication date, retrieval date, title, and URL.

*3) Update Model*: The system updates its analytical model (for examples, see Fig. 5-8) for the domain based on all stored news stories, building analytical data structures for analysis.

*4) Analyze*: The analyst studies analytical data structures and validates the risk model. She exports reports and analytical objects for further study using tools such as Microsoft Excel, graph visualization tools [19], or IEMs [4]. She updates her risk model appropriately, backing unusual or unexpected conclusions with relevant news stories.

*5) Risk Analyzed*: Based on the previous step, new risks are identified and the analyst has analytical objects, reports, and related news stories to support a new risk model.

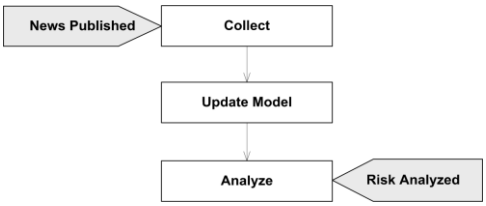


Figure 3. *NewsFerret* Operational Reference Model diagram.

## E. Component Model

The design process of this project decomposed tasks from the above ORM, analyzed functional dependencies, modeled data elements, and structured components based on the non-functional requirements. Details of all those exercises are not presented in this report; a component model summarizes design outputs (see Fig. 4).

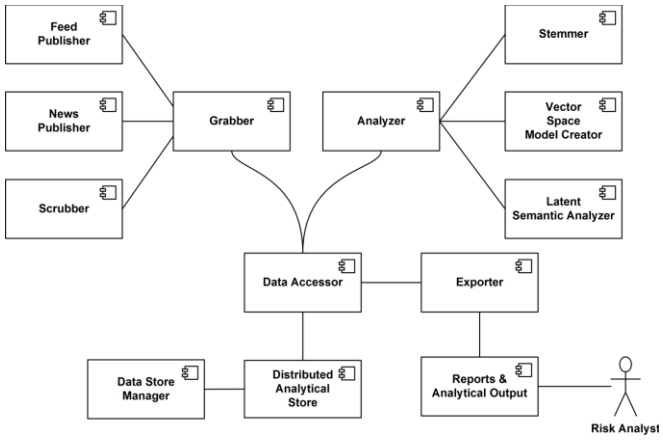


Figure 4. *NewsFerret* Component Model

The diagram and descriptions in Fig. 4 group required functions and data elements to satisfy the functional dependencies and non-functional requirements, illustrating the structure of the *NewsFerret* system from a business standpoint. Component descriptions are listed in operational order and where applicable, notes illustrate the model in the context of the example scenario.

*1) Grabber:* Queries news feeds for news stories related to one or more keyword topics, retrieves the news stories from the News Publisher, and calls the Data Accessor to store metadata and story content. In the example scenario, the component retrieves and stores news stories related to "identity theft" and "credit card theft" topics.

*2) Feed Publisher:* Responds to keyword topic searches, providing a list of recent news story summaries (and URLs), with a typical date coverage range lasting anywhere from within the last 24 hours to within the last week. The Feed Publisher is a publicly available news feed reused for this project. In the example scenario, news feeds return article metadata for news related to "identity theft" and "credit card theft" topics.

*3) News Publisher:* Publishes news stories on the web for retrieval by the News Grabber. This component is provided by news media publishers. In the example scenario, news publishers provide "identity theft" and "credit card theft" news stories in HTML format.

*4) Scrubber:* Extracts the core content of a news story from its published format (typically HTML), removing ad content, menu and navigation content, etc. while retaining the article title, author, and content. We reuse the boilerpipe library for this component [20]. In the example scenario, the scrubber removes ad content and HTML tags from an HTML news article on identity theft.

*5) Data Accessor:* Provides storage-technology-independent data access functions, decoupling the news grabbing and analysis functions from the specific choice of technology for the Distributed Analytical Store. In the example scenario, the data accessor stores article content and metadata for a news story.

*6) Distributed Analytical Store:* Provides storage-technology-specific data access functions for a given distributed data storage technology. Notably, it must store the news stories and the vector space model. This component was realized by reusing the Apache Cassandra database [21].

*7) Data Store Manager:* Storage-specific component that manages the data definitions or schema of the Distributed Analytical Store.

*8) Analyzer:* Coordinates and executes the steps to create the primary analytical objects. Supports scalable distributed processing. The main steps are starting the Vector Space Model Creator to create the vector space model, and then calling the Latent Semantic Analyzer functions to create the semantic space model. In the example scenario, the analyzer performs the above actions against a large set of news stories on "identity theft" and "credit card theft." Resulting semantic space models contain information about related concepts within the news stories. For example, the semantic models show the concept of "password" is very closely related to "amazon" within this set of news stories.

*9) Stemmer:* Reduces words to a stem form to reduce the number of terms with similar meanings. For example, "bank," "banks," and "banking" all reduce to the same stem. Part of this component reuses a publicly available Java implementation of the Porter stemming algorithm [22].

*10) Vector Space Model Creator:* Creates a term-document matrix from the complete set of news stories. A term-document matrix is a matrix where each element (i,j) represents the number of times term i occurs in document j.

*11) Latent Semantic Analyzer:* Implements the key algorithm to create a reduced-rank singular value decomposition (SVD) representing a model of the "semantic space" of all news stories in a threat context. Should operate in a distributed fashion. (Note: this project does not implement distributed LSA processing.) This component reuses several MATLAB [22] built-in functions to realize the algorithm.

*12) Exporter:* Exports reports and analytical objects the risk analyst can use to validate her risk model. (Sample reports and objects are described in the Analysis section.)

*13) Reports and Analytical Output:* Reports and data files used directly by the risk analyst and imported into other tools to validate a risk model. (Example reports and output are provided in the Analysis section.)

### F. Technology Solutions

The *NewsFerret* system implements the above components in Java and MATLAB programming languages. These choices were driven by development familiarity with the languages and availability of open source libraries in these languages to assist with various functions (e.g., scrubbing [20], feed parsing [24], and linear algebra functions [23]).

Due to the potential news story content amount and scalability requirements, core processing and data storage components require a scalable technology, and also must work well together. While in-depth performance characteristics are not evaluated in this report, the system used Apache Hadoop for distributed map-reduce style processing [25] and Apache Cassandra for write-fast, horizontally scalable data storage [21]. This style of storage and processing usage lends flexibility to future *NewsFerret* expansion for gathering large quantities of news stories across a variety of specialized threat contexts and continue providing near-time analysis. The system exports analytical objects in comma-separated value (CSV) or whitespace-delimited format (DAT). This enhances interoperability with other analytical tools such as MATLAB, Gephi (for graph visualization), and spreadsheet applications.

### G. Source Code

The authors of this report published all source code, along with installation, build, and configuration instructions to a publicly accessible GitHub repository [26].

## VI. Analysis

The authors implemented this design and ran the system—like an identity risk analyst would—over several weeks. What follows are results of the completed *NewsFerret* system run and description of how the sample risk model from the example Amazon scenario could be validated and revised based on such results.

### A. Inputs

The first step in using *NewsFerret* is defining a news feed and keyword topic, and thereby defining a threat context, to search for news stories. The authors do not prescribe an exact process, but suggest trial-and-error with some validation. After selective sampling from different news feeds, two news feeds with keyword-based URLs for "identity theft" were selected to represent the example threat context (see Table 2).

TABLE II. News Feeds and Keyword-based URLs for an "Identity Threat" Context

| *Feed Publisher* | *URL* |
|---|---|
| Google News | http://news.google.com/news?um=1&ned=us&hl=en&q=identity+theft&output=rss |
| Huffington Post | http://www.huffingtonpost.com/tag/identity-theft/feed |

### B. Analytical Objects

Next, the analyst performs the necessary steps for creating a set of analytical objects. At any point after news grabbing begins, the analyst runs the Analyzer component and can then export the resulting analytical objects. Although results can be retrieved at any time, these results improved over time as additional stories were gathered. Patterns started emerging and stabilizing after approximately 100 news stories were accrued over approximately 2.5 weeks. The primary analytical output comprises two CSV files: pairwise semantic relatedness measures for important concepts; and top five most related news stories for each pair of important concepts. Fig. 5 illustrates an example snippet of the pairwise relatedness measures, as viewed in a Microsoft Excel spreadsheet.

| | A | B | C | D |
|---|---|---|---|---|
| 1 | source | target | weight | type |
| 251 | amazon | accounts | 0.3285 | undirected |
| 252 | amazon | banking | 0.4195 | undirected |
| 253 | amazon | user | 0.4366 | undirected |
| 254 | amazon | email | 0.4411 | undirected |
| 255 | amazon | unique | 0.4917 | undirected |
| 256 | amazon | shopping | 0.507 | undirected |
| 257 | amazon | website | 0.5671 | undirected |
| 258 | amazon | twitter | 0.6416 | undirected |
| 259 | amazon | group | 0.6474 | undirected |
| 260 | amazon | passwords | 0.6686 | undirected |
| 261 | american | female | 0.2099 | undirected |
| 262 | american | proof | 0.2937 | undirected |
| 263 | american | mortgage | 0.3691 | undirected |
| 264 | american | color | 0.4045 | undirected |
| 265 | american | ebay | 0.4463 | undirected |
| 266 | american | travel | 0.5868 | undirected |
| 267 | apple | ship | 0.2009 | undirected |
| 268 | apple | transunion | 0.2146 | undirected |
| 269 | apple | equifax | 0.2146 | undirected |
| 270 | apple | store | 0.3108 | undirected |
| 271 | apple | express | 0.4302 | undirected |
| 272 | apple | design | 0.8141 | undirected |
| 273 | apple | computers | 0.8494 | undirected |

Figure 5. Example report of relatedness measures for important term pairs. ("Source" and "Target" columns contain two important concepts. "Weight" column contains measure of relatedness between two concepts using a continuous scale between 0 and 1, where 0 means the two concepts are not related at all, and 1 means the concepts are very closely related. "Type" column is for graph visualization tool usage and demonstrates the relations between concepts is not directed. Highlighted cells indicate relations with a high degree of semantic relatedness.)

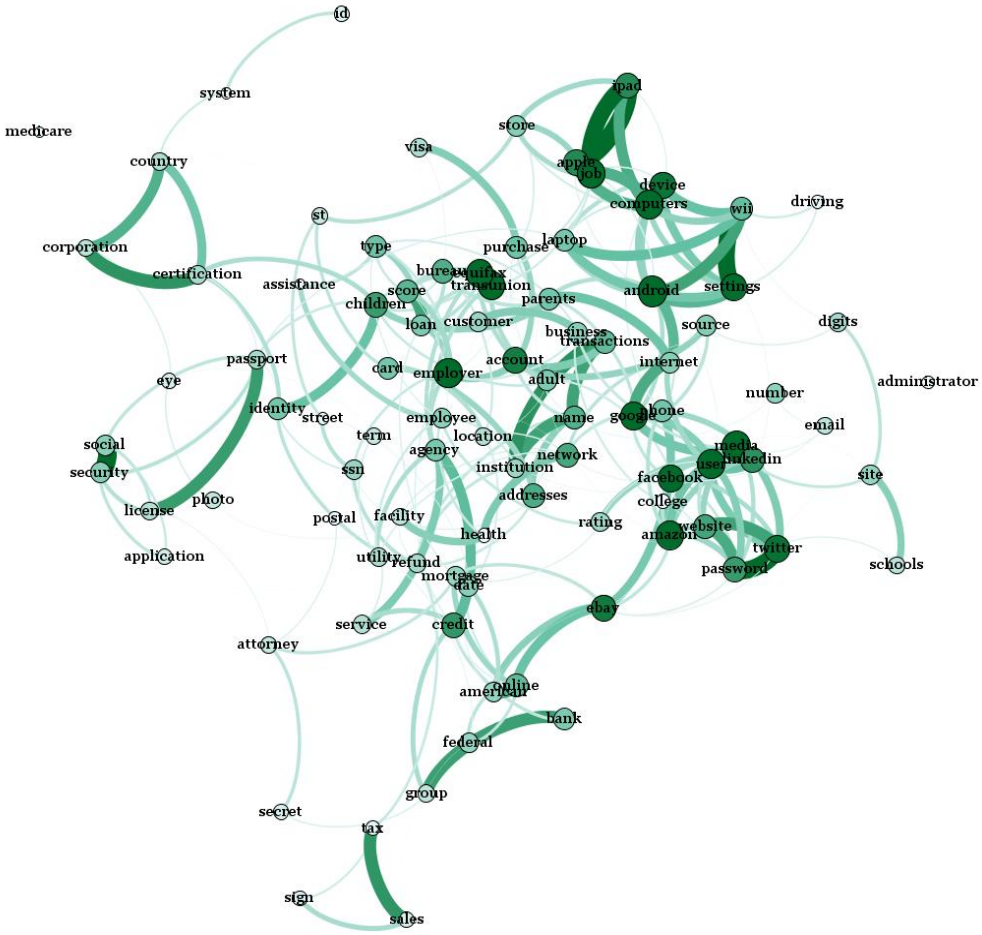


Figure 6. Graph visualization of concept relatedness and importance. (Concepts are represented by labeled nodes; relatedness between concepts is represented by edges. The thicker the edge, the more related the two concepts; the larger the node, the higher the weighted in-degree of the node—a rough measure of the importance of that concept to the identity risk context.

### *C. Concept Relatedness*

Next, relatedness measures are fed into a graph visualization tool for further analysis. The relatedness measures are useful for examining within a spreadsheet, but graph visualization provides a powerful high-level picture of the data. Fig. 6 illustrates a visualization of the same information using the graphing tool Gephi. In exploring Fig. 6, the risk analyst may identify specific relations that are unexpectedly strong. The reason for such relationships is not always clear. For example, "amazon," "facebook," "credit," and "android" display high levels of importance to the identity risk context, while curiously strong relatedness is exhibited between "twitter" and "password."

The graph visualization tool can help future exploration only by showing nodes in direct contact with a specifically selected node of interest (see Fig. 7). The risk analyst must now interpret the analytical models to validate the starting risk model. For example, the relation between "amazon" and "password" concepts could mean that Amazon passwords are being hacked, a large password breach recently occurred, or a host of other issues. This additional tool helps the analyst interpret the above graphs to validate her model.

Furthermore, the relatedness measures shown for pairs of concepts in Fig. 7 indicate the level of meaningfulness in the relationship between two concepts in the context of "identity theft" (the threat context). But what does that mean for any two specific concepts? *NewsFerret*'s analytical models do not state this explicitly or represent cause-and-effect relationships, but do indicate if the two concepts often appear in the same news stories related to the given threat topic.

Accordingly, two techniques are suggested for the risk analyst to analyze and explain this relatedness. First, examining multiple pairs of results offers heuristic guidance in terms of relating concepts. This examination illustrates a pattern. If the relatedness measure is above 0.9, the relationship is likely synonymous or the two concepts are parts of a multi-word term. If the relatedness measure is below 0.2, then there is generally little-to-no meaningful relationship between the two

concepts in this threat context. Typically, relationships of interest to the risk modeler lay between these two extremes.

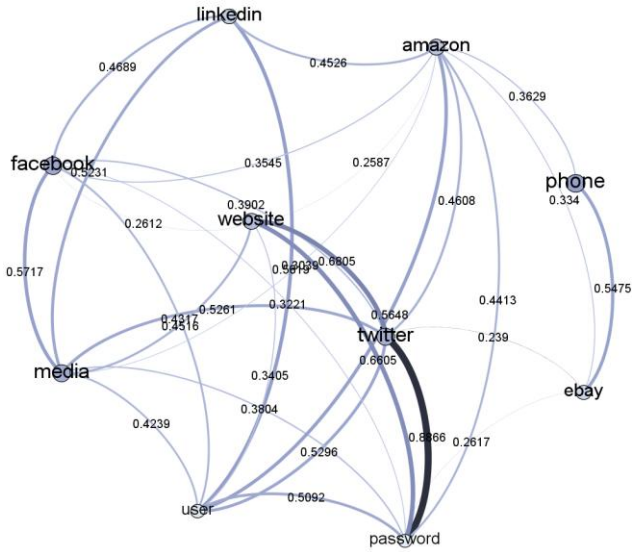


Figure 7. Concepts closely related to Amazon. (All concepts closely related to "amazon" by a measure of 0.2 or higher are shown. Some unexplained relationships between "amazon" and "twitter," "amazon" and "password," "amazon" and "phone" could also be explored.)

However, this heuristic technique is only so helpful. The analyst is still left to explore a large range of related concepts and wants to understand why each of the remaining term pairs is related at certain levels. For example, why do the terms "American" and "travel" receive a high relatedness measure (0.5868) as shown in Fig. 5? The system can repurpose the semantic model as a search engine and conduct a search for the top five most related documents for each concept pair. Table 3 contains a sample of such search results.

TABLE III. ANALYZING CONCEPT RELATEDNESS THROUGH RELATED NEWS STORIES

| *Concept* | *Concept* | *Relatedness* | *Top Related Story Title* |
|---|---|---|---|
| amazon | password | 0.6686 | "Identity Theft -- Your Use of Passwords Could Be Your Only Line of Defense" |
| american | travel | 0.5868 | "U.S. busts massive fake ID scheme" |
| birth | passport | 0.8617 | "Former Lower South man facing identity theft charges here, now in trouble with ..." |
| controls | profile | 0.5721 | "Facebook users risk identity theft, says famous ex-conman" |

By exploring the titles and content of news stories related to the concept pairing, the analyst can find support in explaining why the semantic model concludes two concepts are closely related. For example, based on Table 3, it can be inferred that "amazon" and "password" are related in the risk model because of multiple incidents or alerts about Amazon accounts being compromised by weak passwords. Furthermore, "birth" and "passport" could be related because identity thieves have used falsified birth certificates to acquire passports.

## VII. REVISITING THE EXAMPLE SCENARIO

Equipped with these new analytical models, the following section revisits the example scenario and describes how the analyst could validate and revise the starting risk model. Based on the concept relatedness graph from Fig. 7, the analyst has identified that—somewhat surprisingly—the name, address, and e-mail identity attributes are not as closely related to "amazon" within this threat context as previously theorized. Instead, the "password" identity attribute appears to be at greater risk due to the higher edge weight between "amazon" and "password."

However, edge weight alone is insufficient for making a conclusion. Next, the analyst researches related articles from Table 3. Based on a reading of these news stories, she concludes that "password" is indeed at higher risk. Similar conclusions are reached for "amazon's" relation to "twitter" and "linkedin." Based on these conclusions, the analyst isolates relevant portions of Fig. 7 into a sub-graph (see Fig. 8). Note that "name," "address," and "e-mail" attributes are not represented in Fig. 8 because the edge weight between these attributes and "amazon" was very low (< 0.1).

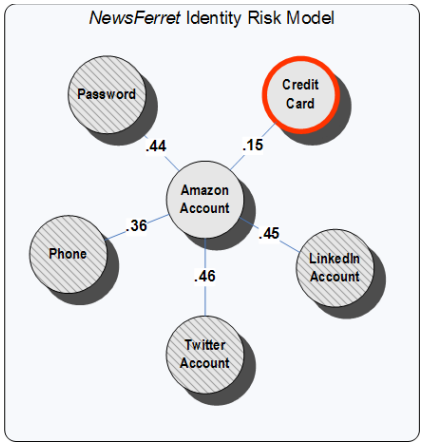


Figure 8. Sub-graph of *NewsFerret* model relevant to example scenario.

Based on Fig. 8, the analyst determines that updates are needed to the original risk model in Fig. 2. After discussion and evaluation with the client, a decision is made to leave in the original attributes of "name," "address," etc., in the model, and add the newly discovered attributes. Fig. 9 illustrates the final risk model. Components with slanted lines indicate identity attributes that have been added to the original risk model based on an interpretation of the *NewsFerret* analytical model.

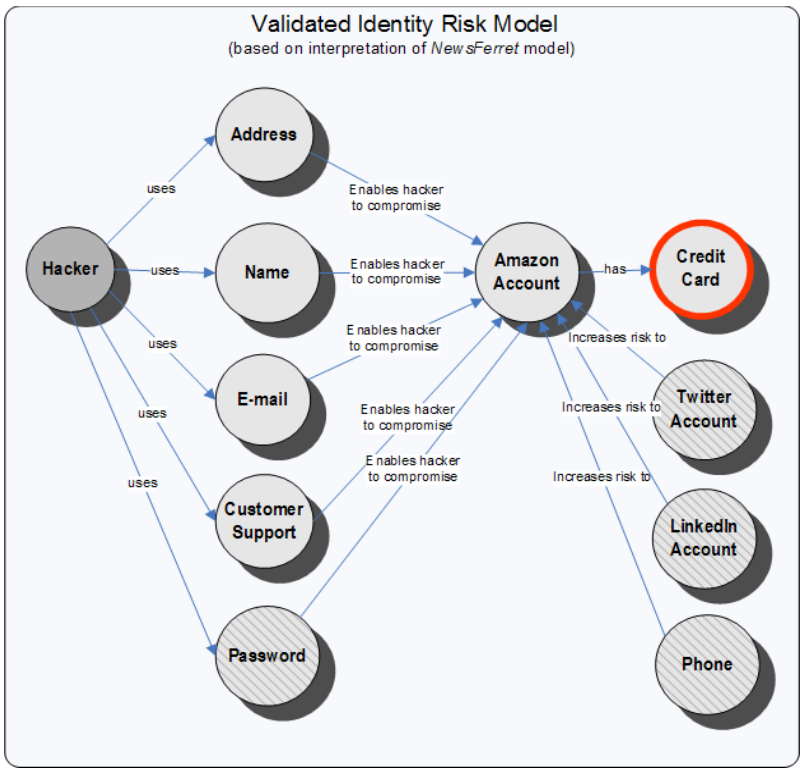


Figure 9. Validated identity risk model for example scenario.

Using the model represented in Fig. 9, the analyst can answer the client's original questions:

- The likeliest attack vectors are "password," "twitter," or "linkedin."
- Missing elements from the original model are "password," "twitter," "linkedin," and "phone."
- Passwords seem to pose the greatest risk. Efforts to improve password practices or authentication techniques should be initiated.
- Users should be cautioned against exposing information on their social media accounts protecting smart phones and Amazon accounts in the event that their phones are stolen.

These recommendations are formulated based on a careful reading and interpretation of relevant news stories returned by *NewsFerret*.

## VIII. Future Research

The authors suggest future work related to this study. Since news articles take around 2.5 weeks to accrue a reasonable level of usefulness, the casual risk analyst would prefer if a set of pre-configured threat contexts were available. The system admin could organize these preset threat contexts by threat type or by another facet like industry vertical (e.g., health care fraud, financial fraud).

Solutions to the one unsolved text mining issue during this study—multi-word concepts were handled somewhat inelegantly—undoubtedly abound. For example "social security number" was treated as three separate concepts, which did not ultimately affect the results; the terms ended up labeling concepts closely related in the graphs. However, this issue could be better handled better by n-gram analysis and perhaps improve interpretability.

In addition, the authors acknowledge that news stories reflect a trendiness bias. Further research could study the modeling bias of news stories on identity threats and other threat contexts to also improve interpretability. Yet despite their problems, news stories have two important qualities benefitting this design: sheer quantity and public nature. First, media outlets publish large numbers of news stories, covering events on thousands of topics from around the world. Over a month-long period, *NewsFerret* found over 200 unique news stories on the keyword topic of "identity theft," even when limited to an English-language U.S. locale. Second, information published in a news story is a form of public record, so in using news stories, the analyst need not be overly concerned with protecting PII or incident details since the information has already been made public.

Moreover, the review of related work noted some differences between security threat modeling and identity risk modeling; the latter must often model a more distributed, inter-connected set of protected resources. While this distinction is useful, the fields of identity and security remain closely related,

and by configuring the right set of keyword topics and a set of security (instead of identity) attributes, *NewsFerret* could be adapted to model and validate cyber-security risks.

Finally, whether a system similar to *NewsFerret* could work with more structured incident reports is an open area for future research. Such a study could confirm if LSA techniques could enable forms of analysis on this data without requiring costly and error-prone manual data entry.

## IX. Conclusion

First, this report outlined the requirements of the identity risk analyst, explaining the need to validate identity, threat, and risk models against the true semantic structure of an identity ecosystem, threat context, and identity risk. The study of this problem is now situated in the context of existing threat modeling research and known text mining approaches for studying semantic structure of documents.

Second, this report illustrates the conceptual framework and functional requirements for a system that can provide some validation of identity risk models. We outline a design capable of satisfying those requirements by gathering and analyzing news stories as representatives of the threat context under analysis. Furthermore, the report introduces and revisits an example identity risk-modeling scenario to demonstrate a sample risk model and its evolution using *NewsFerret*.

We implemented this design and demonstrated the process for configuring and running the system, resulting in analytical output from a month-long run that gathered over 200 unique news stories on the keyword topic of "identity theft." In demonstrating the analytical output and its affect on the example scenario, this report shows that concept relatedness measures can reveal unexpected relations between concepts in a risk model, and that exploring news stories related to the pair of concepts offers further support for those relations.

## References

[1] Javelin Strategy & Research, "2012 Identity Fraud Report: Social Media and Mobile Forming the New Fraud Frontier," Javelin Strategy & Research, Pleasanton, 2012.

[2] Cisco, "2013 Cisco Annual Security Report," 2013. [Online]. https://www.cisco.com/web/offer/gist_ty2_asset/Cisco_2013_ASR.pdf

[3] Kathryn Zickhur, "Three-quarters of smartphone owners use location-based services," Pew Internet, Washington, D.C., 2012. [Online]. http://pewinternet.org/Reports/2012/Location-based-services.aspx

[4] Center for Identity, "Mapping of the Identity Ecosystem," Technical Report TR-010413-2013, 2013.

[5] Center for Identity. (2013) Center for Identity. [Online]. http://identity.utexas.edu/research/projects/identity-threat-assessment

[6] Jason I. Hong, Jennifer D. Ng, Scott Lederer, and James A. Landay, "Privacy risk models for designing privacy-sensitive ubiquitous computing systems," in *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques*, Cambridge, MA, 2004, pp. 91-100.

[7] Shawn Hernan, Scott Lambert, Tomasz Ostwald, and Adam Shostack, *Uncover Security Design Flaws Using The STRIDE Approach*, 2006th ed.: MSDN Magazine, 2006. [Online]. http://msdn.microsoft.com/en-us/magazine/cc163519.aspx

[8] Jerome H. Saltzer and Michael D. Schroeder, "The protection of information in computer systems.," in *Proceedings of the IEEE*, vol. 63.9, 1975, pp. 1278-1308.

[9] Weigo Fan, Linda Wallace, Stephanie Rich, and Zhonju Zhang, "Tapping the Power of Text Mining," *Communications of the ACM*, vol. 49, no. 9, pp. 77-82, September 2006.

[10] Mat Honan, "How Apple and Amazon Security Flaws Led to My Epic Hacking," *Wired*, August 2012.

[11] Lawrence O'Gorman, "Comparing Passwords, Tokens, and Biometrics for User Authentication," *Proceedings of the IEEE*, vol. 91, no. 12, pp. 2021-2040, 2003.

[12] Federal Trade Commission, "Consumer Sentinel Network Data Book for January-December 2011," Federal Trade Commission, 2012. [Online]. http://ftc.gov/sentinel/reports/sentinel-annual-reports/sentinel-cy2011.pdf

[13] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman, "Indexing by Latent Semantic Analysis.," *Journal of the American Society for Information Science*, pp. 391-407, 1990.

[14] Ryan Golden, Ashton Mozano, Yousif Seedham, and Fahd Siddiqui, Mining Capitol Hill Speeches, 2010, Course assignment for EE380L: Data Mining, Spring 2010.

[15] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, *Introduction to Data Mining*.: Pearson Education, Inc., 2006.

[16] Sophia Ananiadou, Douglas Kell, and Jun-ichi Tsujii, "Text mining and its potential applications in systems biology," *Trends in Biotechnology*, vol. 24, no. 12, pp. 571-579, 2006.

[17] Suzanne Barber, Course lecture notes for EE382C: Requirements Engineering, Spring 2011, 2011.

[18] Observer-Dispatch. (2013, March) UTICAOD.com Observer-Dispatch: The Mohawk Valley's Information Source. [Online]. http://www.uticaod.com/news/x1551260940/Rome-woman-charged-with-identity-theft

[19] Gephi. (2013, March) Gephi. [Online]. http://gephi.org/

[20] boilerpipe. (2013, March) boilerpipe: Boilerplate Removal and Fulltext Extraction from HTML pages. [Online]. https://code.google.com/p/boilerpipe/

[21] Eben Hewitt, *Cassandra: The Definitive Guide*. Sebastopol, CA: O'Reilly Media, 2010.

[22] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130-137, 1980.

[23] MathWorks. (2013, March) MATLAB: The Language of Technical Computing. [Online]. http://www.mathworks.com/products/matlab/

[24] Confluence. (2013, March) ROME. [Online]. https://rometools.jira.com/wiki/display/ROME/Home

[25] Apache Hadoop. (2013, March) Apache Hadoop. [Online]. http://hadoop.apache.org/

[26] Ryan Golden. (2013, March) NewsFerret. [Online]. https://github.com/ryancgolden/NewsFerret