The University of Texas at Austin
Center for Identity

The Identity Ecosystem

Razieh Nokhbeh Zaeem
David Liau
Suratna Budalakoti
K. Suzanne Barber

UTCID Report #19-08

July 2019

The Identity Ecosystem

As identity theft, fraud, and abuse continue to grow in terms of both scope and impact, individuals and organizations alike demand a deeper understanding of their vulnerabilities, risks, and resulting consequences. To address this demand, we present the Identity Ecosystem, a novel Bayesian model of Personal, Organizational, and Device Identifiable Information (PII/OII/DII) attributes and their relationships. We populate the Identity Ecosystem model with real-world data from approximately 6,000 reported identity theft and fraud cases. We leverage this populated model to provide unique, research-based insights into the variety of PII/OII/DII, their properties, and how they interact. Informed by the real-world data, we investigate the ecosystem of identifiable information in which criminals compromise PII/OII/DII and misuse them.

We built the Identity Ecosystem into an online tool that answers sophisticated queries. As an example query, it predicts future risk and losses of losing a given set of PII and the liability associated with its fraudulent use. In the Bayesian model, each PII (e.g., Social Security Number) or OII (e.g., Employer Identification Number) or DII (e.g., IP Address) is modeled as a graph node. Probabilistic relationships between these attributes are modeled as graph edges. We leverage this Bayesian Belief Network to approximate the posterior probabilities of the model, assuming the given set of PII attributes is compromised, to answer the query.

Hence, the Identity Ecosystem uncovers the identity attributes most vulnerable to theft, assesses their importance, and determines not only the PII but also the OII and DII most frequently targeted by thieves and fraudsters. The insights the Identity Ecosystem provides are significant, valuable, and sometimes very nonintuitive.

CCS Concepts: • **Mathematics of computing** → **Bayesian computation**; • **Security and privacy** → **Economics of security and privacy**;

Additional Key Words and Phrases: Personally Identifiable Information, Bayesian Computation

ACM Reference Format:
. 2019. The Identity Ecosystem. *Digit. Threat. Res. Pract.* 9, 4, Article 39 (March 2019), 16 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

Identity theft is the signature crime of the digital age [4]. According to the latest U.S. government reports (released in January 2019), it affected an estimated 26 million persons in 2016 [3] and was at or near the top of the Federal Trade Commission's (FTC) national ranking of consumer complaints for close to two decades [6]. Identity theft is a crime that involves the fraudulent acquisition and use of a person, organization, or device's identifiable information (PII/OII/DII).

A person, organization, or device's identifiable information is any data that could potentially identify a particular individual, organization, or device. Many PII, OII, and DII attributes belong to the digital world. The digital world has seamlessly merged into everyday physical world, making a person's identity a complex intermingling of their on-line and off-line attributes. Examples of on-line PII attributes are one's social media accounts, on-line shopping patterns, passwords, and email accounts. Off-line attributes are those related to the physical world such as one's physical

Author's address:

2576-5337/2019/3-ART39 $15.00
https://doi.org/0000001.0000001

characteristics. The same holds for organizations and OII: many OII attributes are digital. When it comes to devices, the majority of the DII attributes belong to the digital space.

While identity theft related crimes continue to be reported with alarming regularity, it is difficult for both individuals and organizations to understand how to react to such events. While, in many cases, the kind of PII compromised is known, people do not know what to do with this information. They also often do not know what information is important and needs to be safeguarded, and what PII might be relatively safe to expose. Similarly, organizations often need to set policies on how to authenticate, what documents to consider for verification, and so on. In a landscape of constantly changing threats, it is difficult for them to come up with policies that trade-off risks and efficiency in a coherent manner. In this paper, we attempt to quantify these trade-offs so that these decisions can be made in a more knowledgeable way.

We have designed and implemented the Identity Ecosystem as a valuable tool that models PII/OII/DII and the ecosystem of identity theft and fraud, analyzes the data, and answers several questions about identity risk and management for both individuals and organizations. For example, the Ecosystem can predict a risk (i.e., probability) of breach for each PII and a potential dollar value damage to the PII owner if the PII is fraudulently used. The risk of exposure of a PII attribute depends on the different methods by which it can be breached. The value of an attribute depends on how it can be fraudulently used, possibly to facilitate further breaches. In addition, once more information about the victim or the incident is available, the Ecosystem is able to refine the predicted risk and value to reflect the new information and converge to the risk and value in the real world.

The Identity Ecosystem stores known data about PII breaches and fraudulent usage in a probabilistic model, and performs Bayesian Network-based inference on this data, to identify high risk and value targets [references withheld for blind review] . Bayesian networks as a statistical tool are a good fit for this problem because of the highly complex interdependence between various attributes. Based on the probabilistic analysis, the Identity Ecosystem tool presents the results as an easy to interpret graph-based visualization, where the attributes are represented as nodes, and the relations are shown as edges. The visualization enables the user to interactively play out various scenarios, and draw conclusions about the risk of exposure and value of the attributes of interest. We present several use cases for the Identity Ecosystem tool throughout the paper. Furthermore, the Identity Ecosystem uncovers and reports statistics of the data: e.g., the percentage of attributes that are isolated from others, top OII attributes used to breach other OII attributes, the distribution of DII attribute risk and values, etc.

In Section 3, we consider use cases which would primarily be of interest to an individual, and use cases which would be applicable to organizations (though there is overlap between the two). Use cases for individuals provide an individual insight into how the information they expose to the world affects them. These use cases could be utilized to guide their decisions on what information to share. They could also help them with quantifying the impact of PII revealed about them, in case of a news report of a breach. The knowledge gained from these use cases could guide them to make informed decisions such as changing PII that could be vulnerable due to a breach, even though they have not been directly revealed. The first three use cases in Section 3 pertain to use cases for individuals. Use cases for organizations can help them set policies around authentication and what documents to request. The last two uses cases in Section 3 are mostly related to organizations. These are only example use cases of Ecosystem and one can take advantage of Ecosystem to answer many more diverse questions.

In this paper, we first briefly review our external source of raw data (Section 2) and then set forth various example use cases of the Ecosystem tool (Section 3). Then, we formally explain the underlying model of the Ecosystem (Section 4) followed by its implementation details (Section 5). We finally conclude and propose future directions (Section 6).

2 BACKGROUND: DATA SOURCES FOR IDENTITY ECOSYSTEM

The Identity Ecosystem provides a framework to investigate identity. However, it needs an external source of information to populate PII/OII/DII attributes and their relationships using real world data. To obtain a list of important PII attributes used throughout this paper, their properties, e.g., initial risk of exposure, and their relationships, we obtained and used the Identity Threat Assessment and Prediction (ITAP) project [9] data.

ITAP [9] is a risk assessment tool that increases fundamental understanding of identity thieves' and fraudsters' processes and patterns. ITAP aggregates data on identity theft from multiple sources (e.g., law enforcement, fraud cases, and news stories) to model and analyze identity vulnerabilities, the value of identity attributes, and their risk of exposure. A team of modelers analyzes identity theft and fraud news stories on a daily basis and models this information using the ITAP schema. For each case of identity theft and/or fraud, ITAP collects and analyzes tools used by criminals, types of information exploited, demographics of victims, etc. In doing so, ITAP captures and analyzes a structured computational model of identity and fraud processes and outcomes [7, 8].

ITAP currently models 5,850 news stories that report on specific identity theft and fraud cases. The current ITAP dataset provides 627 mostly PII (but also sometimes OII/DII) attributes—shown as nodes in the Identity Ecosystem. ITAP also contains the frequency of these PII attributes' appearance as points of entry in breaches and identity theft cases (hereafter called their *risk of exposure*), as well as the intrinsic *values* of the PII attributes (based on the monetary loss of the identity theft cases in which it was misused). Note that when ITAP reports a PII has a high risk of exposure, it means it is highly likely that the identity thieves and fraudsters target and misuse that PII attribute. It does not, however, directly imply that the PII is easily accessible to them. In addition, ITAP currently has 844 connections between PII attributes – shown as directed edges in the Identity Ecosystem – in which a PII was utilized to obtain or fraudulently generate another (in the direction of the edge). In ITAP, weights are associated with these connections based on the frequency of the relationship in identity theft stories [5].

For the record, the ITAP calculates the attribute risk of exposure, the value of an attribute, and the connection edge weight as follows. If A_i is an identity attribute and S_{A_i} is the set of identity theft cases with attribute A_i as an input PII that the fraudsters used as a point of entry, and S is the set of all the identity theft cases in ITAP, then

$$risk = P(A_i) = \frac{|S_{A_i}|}{|S|}$$

$$value = L(A_i) = \frac{\Sigma_{S_{A_i}} MonetaryLoss}{|S_{A_i}|}$$

If T_{A_j} is the set of identity theft cases in which attribute A_j was an output PII, meaning that other PII where used to find out or fraudulently generate attribute A_j, then, using conditional probability (CP)

$$edgeWeight = CP(e_{ij}) = p(A_j\ exposed|A_i\ exposed) = \frac{|S_{A_i} \cap T_{A_j}|}{|S_{A_i}|}$$

3 EXAMPLE USE CASES

The Identity Ecosystem tool can be utilized to investigate and answer many possible identity management questions. In this section, we explain a couple of sample use cases of the Ecosystem.

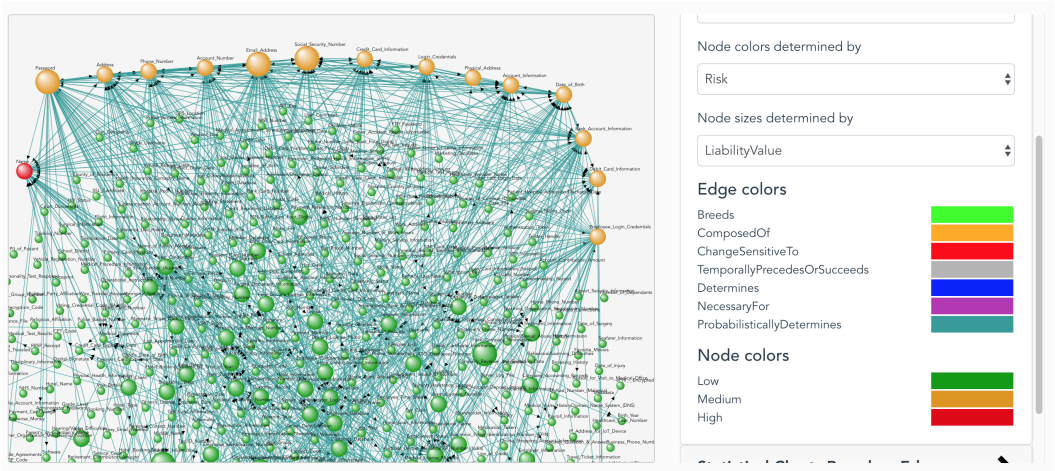

Fig. 1. Risk and Value of PII attributes in the Ecosystem.

3.1 Risk and Values of Attributes

The Identity Ecosystem user can analyze and make decisions about identity risk and value for individuals. The Ecosystem predicts the risk (i.e., probability) of breach for a PII attribute and a potential dollar value damage to the PII owner if the attribute is fraudulently used. PII attributes are connected in various different ways. For instance, low value attributes might be connected to high value attributes; whereby a threat may gain access to a low value attribute, then, via the links present in the connected nature of identity, gain access to other high value attributes. So, low value (small) high risk/highly targeted (red) attributes connected to high value (large) attributes signal trouble.

The Ecosystem Graphical User Interface (GUI) displays PII attributes as nodes and various types of connections between them as edges. The GUI can color and size attribute nodes based on various properties of the attribute, for example, their risk and value. Figure 1 shows the typical set of PII attributes for a person from the ITAP dataset, in which nodes are colored based on their risk of exposure (high risk of exposure in red, medium risk in yellow, and low risk in green) and are sized based on their value (the bigger the node, the higher the dollar value). The Ecosystem user can visually investigate PII attributes, their risk and value, and their connections.

As explained in Section 2, the Identity Ecosystem is populated with real world data of identity theft and fraud cases. The current dataset provides 627 PII attributes shown as nodes and 844 connections between PII attributes shown as edges in Figure 1. The highest risk of exposure, as one would expect, belongs to the “Name” PII (appearing as red which indicates high risk of exposure). While the “Name” of a person is actually a piece of Personally Identifiable Information, it is so widely shared with others and used in identity theft and fraud cases that has the highest probability of exposure. Fifteen other PII attributes have a medium risk of exposure. We can categorize these PII into four classes:

(1) Contact information such as (Physical) Address, Phone Number, and Email Address. It is practically impossible to protect these PII as their sole purpose is to be shared as means of contact.
(2) Banking information such as Account Number, (Bank) Account Information, Credit Card and Debit Card Information. It is not surprising that these PII should be safeguarded.

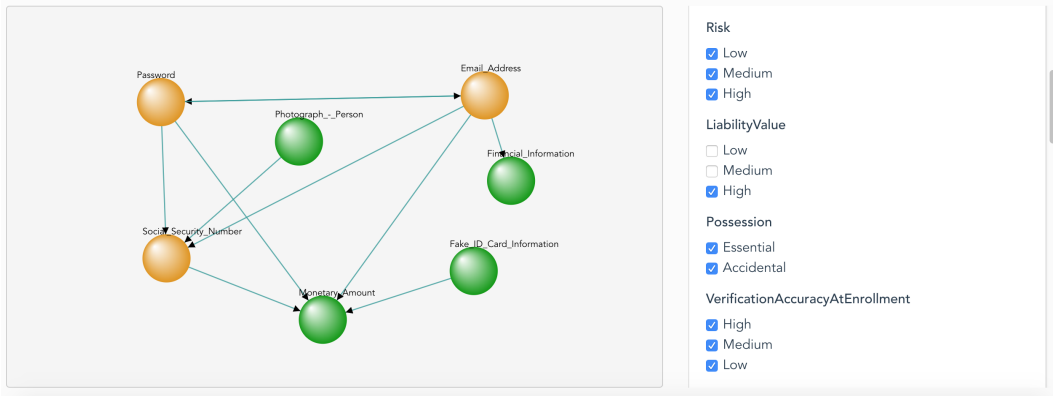

Fig. 2. The connected component of highly valuable PII attributes in the Ecosystem.

(3) PII Frequently used for verification such as Date of Birth and Social Security Number. Since these PII are often used for authentication, it is widely known that they should be safeguarded too.
(4) Digital PII such as Password, Login Credentials, and Employee Login Credentials. We view this category as the most informative. Digital PII are some of the most targeted PII.

All of these medium risk PII are highly connected to other PII.

Top 5% most valuable commonly misused PII as reported by this dataset are as follows:

(1) Contact information: Address, Phone Number, and Email Address.
(2) Verification information: Date of Birth, Social Security Number, and Photograph of a Person.
(3) Digital PII: Login Credentials, Password, and Employee Login Credentials.
(4) Financial information and tax records: W-2 Form Information, and Fraudulent Credit Card Information.

It is interesting that medical information do not directly show up in this list of high value PII. Out of the above high value PII, Social Security Number, Email Address, and Password are at the medium risk level. Figure 2 displays the *connected component* of highly valuable PII.

High value OII as reported by this dataset are Company Identifying Information and Organization Proprietary Information.

3.2 High Level Understanding of Identity

The Ecosystem distinguishes various properties of identity attributes, such as the attribute's type. The attribute's type for a person is divided into four type categories:

- What You Are: a person's physical characteristics, such as fingerprints.
- What You Have: credentials and numbers assigned to a person by other organizations, such as Social Security Card.
- What You Know: information known privately to a person, such as passwords.
- What You Do: a person's behavior and action patterns, such as GPS location.

The Ecosystem tool can highlight attributes of each type, or provide combinations, to answer questions like "what are the most valuable credentials a person owns?"

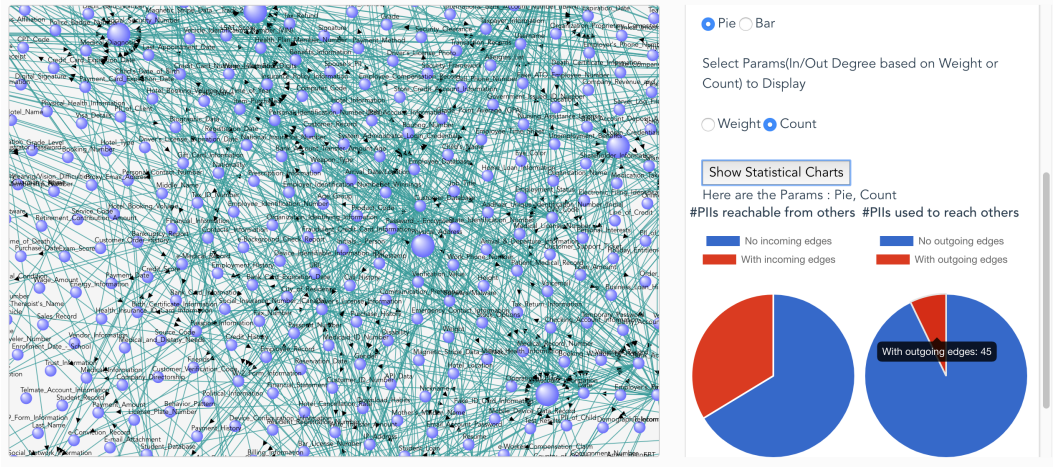

Fig. 3. Pie Charts of InDegrees and OutDegrees of PII Attributes.

3.3 Statistics of Identity Ecosystem

The Identity Ecosystem is capable of displaying a variety of statistical analytics. For example, Figure 3 shows the number of PII attributes that can be used to reach others ($outdegree > 0$) and the number of PII attributes that are reachable from other PII ($indegree > 0$). Notably, in this dataset, 65% of the PII are disconnected from the rest of the graph, which means that, even though they have been misused, there is not a strong probability that indicates they reveal or are revealed by other PII. It is also interesting that, in over 600 PII, only 45 have outgoing edges. This means that these 45 nodes are the common starting points for identity fraud. We classify these 45 nodes that should be particularly protected as follows:

(1) Contact information such as Name, (Physical) Address, Phone Number, and Email Address. As noted before, one cannot fully withhold this class of PII as they are means of contact. However, as evident in data breaches, they are particularly useful to identity fraudsters.
(2) Banking information such as Account Number, (Bank) Account Information, Credit Card and Debit Card Information, CVV Code, Expiration Date, Routing Number, PIN number, and Check Information. These PII have been gateways to other financial information in identity theft and fraud cases.
(3) PII Frequently used for verification such as Date of Birth, Social Security Number, Birth Certificate Information, Driver's License Number, Passport Information, Signature, and Photograph of a Person.
(4) Digital PII such as Username, Password, Login Credentials, and Employee Login Credentials.
(5) Medical information, for example, Insurance Policy Information and Patient Medical Record.
(6) Tax records, e.g., W-2 Form Information and Employee Record.
(7) Biographic Data.

3.4 Authentication by Organizations

Authentication, from the perspective of an organization, is a method for verifying that a person is who he/she claims to be, so that a resource is being accessed only by persons who have a legitimate claim to it. Usually a set of attributes is used during the enrollment process, and another (often smaller) set is used to authenticate the person after enrollment. The authentication process

exposes the organization to an obvious risk: it may be possible for someone to falsely authenticate themselves and gain access to privileged data. Another risk is that the enrollment information stored by the organization exposes it to potential liability if it is accessed illegally. For this reason, it is valuable for organizations to know both the likelihood of a false authentication given the data they use for the process, and the future exposure risk and liability they may be exposed to in case of a breach.

The Ecosystem assigns different properties to PII attributes, among which *Accuracy at Enrollment* measures how accurately an attribute can be verified in the authentication enrollment phase. To reduce an organization's liability while also increasing their authentication accuracy, it is best to use a low risk, low value attribute that provides high accuracy at enrollment. We have leveraged the Identity Ecosystem to investigate the problem of finding an optimal set of PII–that satisfy authentication purposes but minimize risk of exposure–elsewhere [references withheld for blind review] .

3.5 The Breeding Relationship

To breed a PII attribute or document is to create a (legitimate or fraudulent/counterfeit) instance of it. For example, a Birth Certificate can be utilized to create—or breed—a legitimate Passport. The Ecosystem shows various relationships (edges) between PII attributes, including the breeding relationship. The Ecosystem can determine, given an attribute for a person, the probability that other attributes can be fraudulently bred. The Ecosystem user can focus on the first, second, or more order of connectedness for PII attributes to go through multiple steps of breeding PII attributes or documents.

3.6 Ecosystem Queries

The Ecosystem is capable of answering non-trivial questions relevant to the overall risk and liability of any person or organization in terms of managing identity attributes. For instance:

(1) Effect of exposure: When a set of attributes is exposed, how does it affect the risk of other attributes being exposed? For instance, if the social security number (SSN) of an individual is compromised, what are the most risky PII items that fraudsters might try to obtain after that? To answer this question, the user can run the query *Infer probability of breach based on evidence*, in the new window that opens choose *social security number (SSN)* as evidence, and run the query. Using the Bayesian inference, the Ecosystem calculates the change in the probability of exposure after the compromise of SSN, which is reflected by the change of color of the nodes. The Ecosystem also shows the *predicted expected loss* because of the SSN compromise.
   For this particular example, after the exposure of SSN, the risk of exposure for 111 PII is affected. The following PII have the highest risk of exposure after the incident:
   (a) From contact information: Name, Address, Phone Number, and Email Address.
   (b) From banking information: Credit Card Number, Debit Card Information, Bank Card Expiration Date, CVV Code, and Check Information.
   (c) From PII frequently used for verification: Date of Birth, Driver's License Information, and Passport Information.
   (d) From digital PII: Username, and User Credentials.
   (e) From medical information: Insurance Policy Information, and Patient Database.
   (f) From tax records: Employee Record.
   One can compare the effect of this breach with the breach of Driver's License Information that affects only 65 PII where the most at risk PII after the exposure are Name, [Physical] Address,

Social Security Number, ID Card Information, User Credentials, Patient Medical Record, and Employee Record. As evident from these two predictions for breaches, the exposure of SSN has greater consequences. Even though the exposure of Driver's License may in turn expose SSN, chances are lower as an edge with a probability less than one goes from SSN to Driver's License.

(2) Cause: If a set of attributes have been exposed, what was the most likely origin of the breach? As an example, if an individual finds out that his/her credit card number is compromised, the Ecosystem can help to *Detect most probable origin of breach* through selecting Credit Card Number as the evidence and running the query. Based on the ITAP data, the likely origins (in the order of likelihood) in the event of this breach are User Credentials, Expiration Date[1], CVV Code, Customer Database, Username, Email Address, Password, Employee Login Credentials, and Login Credentials. All of these reported likely origins are either directly related to credit cards, or are digital.

(3) Cost/Liability: What is the total cost/liability of an attribute being exposed, in terms of increased risk of exposure of other attributes? Which attributes have the highest liability (the product of the expected cost and the probability of exposure) and are breach hot-spots that should be best protected? Ecosystem can answer this query through *Find breach hot-spots* and reviving the evidence of the breach. For instance, after the breach of Social Security Number, the top five hot-spots are Password with liability over $20K, Photograph of a Person with over $14K, Date of Birth with over $4K, W-2 Form Information with over $2K and Debit Card Information with over $1K.

Note that these liability numbers are not per victim, but are deducted from per identity theft and breach *case* data, which explains why they are so high. Consequently, these numbers are very useful to companies, organizations, or other holders of PII that are breached. The liability value indicates the amount of money recommended to spend to protect a set of PII that the company holds, after the company is breached and is aware of the exposure of a PII.

4 ECOSYSTEM MATHEMATICAL MODEL

In this section, we elaborate on the mathematical model behind the Identity Ecosystem.

4.1 Modeling Identity Attributes and Relationships

We define a person, organization or device's identity as a set of informational data that are linked to the person, organization, or device. Each such piece of information is called an attribute. *Name*, *age*, *zip code*, and *Social Security Number* are examples of such attributes for a person. *Federal Tax ID*, *Owner*, *Vendor Number*, and *Website URL* are examples of such attributes for an organization. For a device, we can mention *IP Addresses*, *Manufacturer*, *Operating System Type*, and *Serial Number* as examples.

Attributes can be classified in many ways depending on their properties, such as, whether or not an attribute is unique to an entity, whether or not an attribute is widely used, how accurately it can be verified, etc. For example, attributes like name or zip code are applicable to any person but are not unique to a person and cannot be used on their own for verification or authentication purposes. On the other hand, SSN is unique to a specific person and hence a very good candidate for authentication.

We identify several different properties for attributes:

[1]It may seem counterintuitive to state the Expiration Date, or CVV Code, as the origin of the exposure of a Credit Card Number. However, note that ITAP identity theft and fraud cases report multiple PII as inputs and outputs of a theft/fraud. As a result, the Expiration Code or CVV are involved in the exposure of Credit Card Number, but may not be the sole origin.

(1) Type (categories PII based on their nature): What You Are, What You Have, What You Know, What You Do.
(2) Risk (shows the risk of exposure): Low, Medium, High.
(3) Liability Value (shows the monetary loss to the individual or organization if compromised): Low, Medium, High.
(4) Possession (identifies if individuals/organizations/devices necessarily have the PII/OII/DII): Essential, Accidental.
(5) Verification Accuracy At Enrollment (measures how accurate it is to verify one's PII/OII/DII at enrollment): Low, Medium, High.
(6) Prevalence (shows what percentage of the population have the PII/OII/DII): Ubiquitous, Common, Rare.
(7) Uniqueness (shows how unique the PII/OII/DII is for the individuals/organizations/devices that have it): Individual, Small Group, Large Group.
(8) Verification Invasiveness (shows how invasive it is to verify one's PII/OII/DII): Low, Medium, High.

The Ecosystem displays each attribute as a node. It can color or size the attributes based on their properties. Once the user selects a property on which to base the color or size, all the identity attribute nodes will be colored or sized based on their current value of the selected property, as in Figures 1 (colored based on risk and sized based on liability value).

Identity attributes are related to each other in many different ways. For example, one attribute can determine another, one attribute can be used to generate another, or one attribute might be composed of many other attributes. We recognize the following relationships between identity attributes α and β:

(1) α Breeds β means that an instance/value of α may be used in order to create a legitimate or fraudulent instance/value of β. For example, driver's license breeds many other documents like boarding pass.
(2) α Composed Of β means that for any value α_i of the attribute α there is a value β_j of the attribute β such that β_j is a proper part of α_i. For example, full name is composed of first and last names.
(3) α Changes Sensitive To β means that for any person P with attributes α and β, if the value of β changes for P, then the value of α changes for P. For example, one's photograph changes with age or one's driver's license changes with address.
(4) α Temporally Precedes β means that for any person P, P must possess some value of attribute α before P can possess a value of attribute β. For example, a student identification number temporally precedes a degree.
(5) α Determines β means that for any person P with attributes α and β, the value of α possessed by P implies the value of β possessed by P. For example, date of birth determines age.
(6) α Necessary For β means that for any person P, if P has a value for the attribute β, then P has a value for the attribute α. For example, a passport number is necessary for a passport.
(7) α Probabilistically Determines β means that for any person P with attributes α and β, P's having a given value of α implies that P probably has some particular value of β. For example, one shares her spouse's last name with a probability, therefore one's spouse's last name probabilistically determines one's last name.

The relationship between two attributes α and β is shown with a directed edge from α to β in the Ecosystem. The user can select to view one or multiple types of edges at a time.

The ITAP dataset currently supports only the *probabilistically determines* type of relationships. However, we have created a manual set of other types of relationships between PII/OII/DII.

4.2 Modeling Identity Ecosystem

We represent the Identity Ecosystem as a graph $G(V, E)$ consisting of N attributes (nodes) $A_1, ..., A_N$ and a set of directed edges between pairs of nodes. Each edge $e \in E$ is represented as a tuple $e_{ij} =< i, j >$ where A_i is the originating node and A_j is the target node such that $1 \leq i, j \leq N$.

We define the set of all incoming edges to A_j as $IN(A_j) = \{e_{xy} | e \in E \wedge y = j\}$, and let the set of all parents of A_j be $PARENT(A_j) = \{A_x | e_{xj} \in E\}$.

Each node A_j is labeled with a Boolean random variable, denoted $D(A_j)$, which is *true* if the attribute has been exposed/breached and *false* otherwise. Each edge e_{ij} represents a possible path by which A_j can be breached given that A_i is breached[2]. For simplicity, we consider all edges to be independent[3]. Therefore, we can assign conditional probabilities to each edge $CP(e_{ij}) = p(D(A_j)|D(A_i))$.

Consequently, the Identity Ecosystem model consists of:

(1) a set of nodes V, each corresponding to an attribute,
(2) a set of edges E, such that a directed edge exists between any nodes A_i and A_j if and only if A_i impacts the risk of exposure of A_j, and
(3) a list of conditional probability estimates for each node A_j, representing how the parent nodes $PARENT(A_j)$, impact the risk of the child A_j.

Also as part of the model, each node has a prior probability $P(A_i)$ (obtained from ITAP) of it getting exposed on its own (as the first breach in the network). For example, a person's date of birth has a higher prior probability of being exposed than his/her driver's license number, because people are less careful with the former than the latter, and the date of birth appears more frequently as an entry point in the ITAP identity theft cases.

We also assume that each node has a monetary loss value, $L(A_i)$ (from ITAP), which represents the amount a person/organization loses (intrinsically) in case the corresponding attribute is exposed. This loss does not include any secondary loss. For example, the exposure of one's driver's license information incurs relatively lower intrinsic cost, even though there might be scenarios where it may lead to further losses in future, by leading to the exposure of other more sensitive data. The model only assumes that the intrinsic loss value is provided. (In this example, the intrinsic loss value of driver's license information is about $400K. However, the Ecosystem reports, through the execution of the third query, that the further losses incurred by the exposure of drivers's license information is more than $4.4M, as it affects the chance of exposure for multiple hot-spot PII, particularly the top five hot-spots after this breach: Social Security Number, Password, Date of Birth, Phone Number, and Debit Card Information.)

4.3 Background: Bayesian Network-based Inference

The formal framework we defined above has significant similarity to machine learning tools broadly known as graphical models. Graphical models allow us to represent a complex network of probabilistically dependent or correlated random variables and perform inference on the model.

Bayesian networks are a probabilistic graphical model-based approach that can very effectively represent probabilistic causal dependencies in a state space. For any set of N k-variate discrete random variables with inter-dependencies, the joint probability distribution would need to tabulate a total of k^N possible states. However, in many practical situations with a large set of random

[2]The probabilistically determines edges provided by the ITAP dataset directly suit this purpose. Other types of edges, too, imply the breach of one attribute based on the breach of the other with a certain probability.

[3]In a more general setting where edges are not necessarily independent, e.g., where multiple attributes are needed to breed a new attribute, we define joint probability distributions on all the edges in $IN(A_j)$ for each node A_j. The joint probability distribution can be defined as function of all $D(A_i)$ such that A_i belongs to $PARENT(A_j)$.

variables, many variables are independent of each other or the causality is indirect. Bayesian networks take advantage of this by only requiring the representation of direct causal dependencies.

Visually, a Bayesian network can be represented as a directed graph. Random variables are represented by nodes, while directed edges are used to represent causal dependencies, with the direction of the edge from the causal to the impacted variable. The causal variables for any node are usually referred to as its parents. The probabilistic dependence of a node on its parents is represented via a conditional probability distribution (CPD). So, if a node has m parents, and each of which can take k states, a CPD specifying the state probabilities of the child node for each of the k^m combinations of parents' states would need to be described. For a graph with N nodes, approximately Nk^m conditional probability values are needed. However, this value is still much smaller than k^N values that would need to be stated in the naive case of not considering parents.

Even in the absence of information about the current state of any node in a Bayesian network, priori probability estimates can be calculated in principle for each node via marginalization. In case evidence becomes available that a certain subset of variables has taken a certain value, new probability distributions incorporating this new information can be calculated for the other nodes. However, since a naive marginalization approach is usually computationally prohibitive in practice, more efficient algorithms have been developed. We use a variation of a belief propagation algorithm, the Junction Tree algorithm [2], for the Identity Ecosystem.

4.4 Mapping the Identity Ecosystem to a Bayesian Network

The Identity Ecosystem model consists of a set of nodes V with edges between them E, and a list of conditional probability estimates, providing for a set of known cases, the probability of a child node being exposed, given that a particular subset of one or more parent nodes have been exposed. Formally, for a node A_i, such a list consists of statements asserting $p(D(A_i)|D(R)) = m$, where m is a probability value ($0 \leq m \leq 1$) and $R \subset PARENT(A_i)$. Note that the set of known cases is not necessarily comprehensive, i.e., there might be subsets of $PARENT(A_i)$ for which p is not given. Another way to provide this list is that for each node A_i with parent set $PARENT(A_i)$, a probability function $f(T_i, A_i) \rightarrow [0, 1]$ must be provided, where $T_i \subset P(PARENT(A_i))$ (i.e., T_i is a subset of the power set of $PARENT(A_i)$).

To complete the Bayesian network model, for each node A_i, this list of probability estimates should be converted to a conditional probability distribution (CPD) table. That is, a probability value must be assigned to all possible combinations in which parent attributes can become known. To generate such a complete CPD for a node, we need to construct a function $g(C_i, A_i) \rightarrow [0, 1]$, where $C_i = P(PARENT(A_i))$, the power set of the set of parents of A_i.

For a node with k parents in the Identity Ecosystem, $2 \times 2^k = 2^{k+1}$ is the number of values we need to specify for A_i. Assigning this probability value is straightforward for members of C_i also present in T_i, as these are already known. For a combination $K \in C_i$ not present in T_i, the probability that a malicious entity successfully exposes an attributes is 1 minus the probability that it fails to do so after trying every applicable member of a set X, such that $X \subset T_i$ and all attributes that are part of any member of X, are also members of K. We prune this set a bit further, making the assumption that, if an exposure attempt using more information will fail, an exposure attempt using a subset of that information will fail as well. In practice this means that, all members of X that are a subset of another member of X, are removed from the set. We call this set the input set, writing it as $I(K \in C_i, T_i)$.

Thus formally, for a node A_i such that C_i is the power set of its parents, for any $X \in C_i$, $g(X, A_i) = f(X, A_i)$, if $X \in T_i$, else $g(X, A_i) = 1 - \Pi_{Y \in I(X,T_i)} f(Y, A_i)$. This gives us a recursive definition for calculating the conditional probability distribution for a node, given a list of conditional probability estimates.

Table 1. List of identity attributes in the USCIS N-400 form.

Data required by USCIS			
1. MilitaryId	2. MilitaryServiceRecord	3. Email	4. SSN
5. PhoneNumber	6. TravelHistory	7. Fingerprints	8. ParentsName
9. ParentsOccupation	10. SpouseInfo	11. Address	12. BirthCertificate
13. Hometown	14. School	15. Organization	16. Signature
17. Citizenship	18. ZipCode	19. Name	20. DateofBirth
21. Height	22. Weight	23. Gender	24. EyeColor
25. HairColor	26. Ethnicity	27. CrimeHistory	28. Age

4.5 Using Bayesian Network to Answer Queries

The Ecosystem tool, in its present form, is capable of using Bayesian inference to perform three chief kinds of analysis (as explained in a use case in Section 3.6): 1) analyzing the risk of exposure, 2) inferring the most likely source of a breach, and 3) calculating the expected cost of attributes.

4.5.1 Analyzing the Risk of Exposure. In this analysis, we are interested in finding out the effect of a breach. In other words, given that a set of attributes $BREACHED = \{A_i | A_i \in V \wedge D(A_i) = true\}$ have been exposed/breached, we want to know: a) for any attribute $A_j \notin BREACHED$, the expected/conditional probability $P'(A_j)$ of breach given the evidence that the attributes in $BREACHED$ have been exposed, and b) the expected increase in cost/liability C due to the breach.

We model this query as an inference problem by treating the breach as input evidence. We set the exposure evidence values $D(A_i)$ for the nodes in $BREACHED$ to *true*, and use Bayesian inference to compute the posterior probabilities for each node in the system. Note that these posterior probabilities are exactly the $P'(A_j)$ values that we need. Now, given the $P'(A_j)$ values, it is easy to compute the percentage increase in the risk and compute an expected increase in cost/liability as $(P'(A_j) - P(A_j)) \times L(A_j)$. Hence, the total cost of the breach can be computed as $C = \Sigma_j (P'(A_j) - P(A_j)) \times L(A_j)$. Note that $P'(A_i) = 1$ if $A_i \in BREACHED$.

Section 3.6 gave examples of this query. As another example, consider the breach of one's credit card information. The highest risk of exposure ($P'(A_j)$) after this breach belongs to Bank Account Information followed by Name, Date of Birth, Social Security Number, W-2 Form Information, User Credentials, Driver's License Information, Signature, Employee Record, Check Information, Birth Certificate Information, Insurance Policy Information, and Patient Database.

One can surely use Ecosystem to find the effect of exposure of personal information in real world scenarios. We offer a practical example here: consider the PII required when applying for Naturalization at U.S. Citizenship and Immigration Services (USCIS). Required PII indicated in the N-400 form[4] are shown in Table 1. If an applicant's form falls in the wrong hands and all the information of Table 1 get breached then the most at risk PII after this breach are: Username, Debit Card Information, W-2 Form Information, Personal Identification Number (PIN), Biographic Data, Passport Information, Check Information, Routing Number, Expiration Date, and Patient Database. Note that while many PII are shared between the responses to the first query, the degree of the effect of exposure and certain PII vary.

4.5.2 Inferring the Most Likely Source of a Breach. The next problem we address it that of finding out the source of a breach (or a set of breaches). In other words, given that a set of attributes

[4] "U.S. Citizenship and Immigration Services", https://www.uscis.gov/ (accessed May, 2018).

$BREACHED = \{A_i | A_i \in V \wedge D(A_i) = true\}$ have been exposed/breached, we want to know the most probable source of the breach/breaches.

Again, this query can be modeled as an inference problem in a Bayesian Network. We set the evidence of exposure for the nodes in *BREACHED* to *true* (i.e., set $D(A_i) = true$) and then compute the posterior probabilities for all nodes in the system using the Junction Tree algorithm [2]. During this calculation, we focus only on the ancestor nodes of the nodes that were breached. An ancestor node of a node A_i is any node A_j such that a directed path lies between A_j and A_i, i.e., there is a chain of causality from A_j to A_i via which A_j can be responsible for a breach of A_i. In addition, the evidence of breach information ($D(A_i)$) for all the child nodes of the breached nodes are set to zero. This assures that the updated posterior probability estimates of all nodes follow the causal direction, in case of cycles.

Thus, for each ancestor node $A_j \in ANCESTORS(A_i)$ we compute the posterior probability $P'(A_j)$. If $P'(A_j) - P(A_j) > 0$, then A_j is a possible source of the breach. To discover the most likely original source of a breach, we first select the highest common ancestors among the possible sources and then report the one with the highest increase in the posterior probability.

For an example of the use of this query, Ecosystem reveals that the most probable origins of a breach of Driver's License Number are Driver's License Information (i.e., card), Password, Employee Login Credentials, Login Credentials, Email Address, Bank Account Number, Routing Number, User Credentials, Account Number, and Personal Identification Number (PIN), in order.

4.5.3 Expected Cost of Attributes. For this analytic, we are interested in finding out the cost/liability of managing an attribute. In other words, we would like to find out how an attribute's exposure increases the risk of other attributes' getting exposed, so that an attribute incurs not only its own intrinsic cost, but also some expected costs downstream.

One might be tempted to simply use the same idea as the first analytic where we analyzed the effect of a known breach. But this situation is different. Now we want to analyze all the possible scenarios where a specific attribute A_i is part of a breached set. In other words, there is no unique set to start with. Also note that the overall cost will not be the same with a different breach set. Another aspect of the problem is that, given a specific breach set, it is possible to compute the total expected cost of the breach using the same techniques as before, but it does not say how the increased cost should be apportioned between the sources of the breach. To take into account all these factors, we propose the following solution.

Note that the problem of distributing the increased cost of exposure arises because there are multiple parents for each node and any subset of them could have been in the breach set. So, instead of trying to sample all possible breach sets, we simply reverse the question and ask, given that an attribute A_j is exposed, what are the expected sources and how much is each of them responsible. Let, $ANCESTORS(A_j)$ be the set of ancestors of A_j. We compute the posterior probability $P'(A_k)$ of exposure for each attribute $\{A_k \in ANCESTORS(A_j)\}$, after setting the belief that A_j is exposed. Note that since A_j has its own intrinsic risk $P(A_j)$ of being exposed and an intrinsic cost $L(A_j)$, the liability attributed to all the ancestors would be $L_j = (1 - P(A_j)) \times L(A_j)$. Now we define the liability attributed to an ancestor A_k for causing the breach of A_j as $L_{kj} = (P'(A_k) - P(A_k)) \times L_j / S$, where $S = \Sigma_{\{l|A_l \in ANCESTORS(A_j)\}}(P'(A_l) - P(A_l))$. Finally, we define the total cost/liability of an attribute A_i as $C = P(A_i) \times L(A_i) + \Sigma_{j \neq i} L_{ij}$.

For example, consider two scenarios: the breach of Credit Card Number and the breach of Driver's License Number. In the former case, the top five hot-spots and their liability values are Social Security Number with the liability of $16,746,485, Email Address with $15,998,216, Photograph of a Person with $6,728,550, Address with $5,494,300, and Employee Login Credentials with $4,573,032. In the latter case, the top five hot-spots are Social Security Number with the liability of $24,608,921,

Password with $10,120,835, Date of Birth with $4,027,446, Phone Number with $3,400,008, and Debit Card Information with $1,867,538.

Finally, note that it is possible to run each of the three queries with multiple evidence. It is further possible to run consecutive queries to monitor the effect of multiple exposures over time. Table 2 displays some PII and the answer to each of the three queries for them as evidence. Some of the findings shown in this table are as follows:

(1) For some critical PII such as Birth Certificate and W-2 Form Information, likely sources of origin are few and clear. Two important origins of breach in these cases are Driver's Lisence Information and digital login credentials (e.g., Email Address and Employee Login Credentials).
(2) Social Security Number consistently ranks high with high liability in the hot-spots.
(3) It may seem surprising to list Name, or Address, as the most at risk PII after a breach, since these PII are usually not even protected for a single individual. Note that, however, ITAP identity theft and fraud cases concern organizations that keep customer data too, which means that the Names and Addresses of *all* the customers should be protected after a breach.

5 ECOSYSTEM IMPLEMENTATION

In this section, we review our current implementation of the Identity Ecosystem.

5.1 Graphical User Interface

The Ecosystem GUI (see Figure 1 as an example) consists of two panels: the main panel, and the right panel. The main panel renders the graph of attribute nodes and edges. The right panel includes Filters, Display Controls, and Statistical Charts.

- The Filters sub-panel provides the capability to filter the Ecosystem nodes and edges and it has two parts: Node filters and Edge filters.
  - Under Node filters, the user can choose to show nodes of a certain value for a certain property, e.g., nodes with the *essential* value for the *possession* property.
  - Under Edge filters, the user can choose to display no edges, all edges, or only some types of edges, e.g., the *determines* and *probabilistically determines* edges only.
- The Display sub-panel is dedicated to general controls of the GUI such as showing/hiding node names.
  - The user can also choose to display a subtree of a selected depth rooted at a selected node (Figure 4), or find the shortest path between two nodes.
  - The Color/Size options allows the user to color or size the nodes based on any of the node properties, and view the index of Edge Colors and Node Colors.
- The statistical charts are based on Nodes or Edges.

5.2 Implementation Details

The Identity Ecosystem tool was implemented in JavaScript with Vue.js. We used D3 [1] for visualization.

6 CONCLUSION

This paper presents a graph-based model to represent the relationships between various personal, organizational, and device identifiable information attributes. We mapped this model, the Identity Ecosystem, to a Bayesian network. Provided with actual identity theft and fraud input data from about 6,000 identity theft and fraud cases for the first degree probabilities of the Bayesian network, the model enables sophisticated inference that can answer many interesting questions in the identity

Table 2. Answers to Queries with Various Evidence.

Evidence	Q1: Top 5 At Risk PII After Breach of Evidence	Q2: Top 5 Likely Source of Breach	Q3: Top 5 Hot-Spots and Their Approximate Liability
Email Address	Name Credit Card Information Address Date of Birth Social Security Number	Address User Credentials Login Credentials Passport Information Employee Login Credentials	Social Security Number ($27M) Address ($6M) Financial Information ($5M) Login Credentials ($5M) Date of Birth ($4M)
Date of Birth	Password Name Address Social Security Number Phone Number	Bank Account Number Social Security Number Routing Number Account Number Name	Password ($34M) Social Security Number ($27M) Email Address ($9M) ID Card Information ($7M) Phone Number ($4)
Birth Certificate	Username Name Address Date of Birth Social Security Number	Driver's License Information Employee Login Credentials	Social Security Number ($27M) Password ($20M) Photograph of a Person ($4M) Date of Birth ($4M) Phone Number ($4M)
Username Password	Bank Account Information Name Address Date of Birth Social Security Number	Bank Account Number Date Bank Account Information Phone Number Address	Social Security Number ($25M) Photograph of a Person ($7M) Login Credentials ($4M) Date of Birth ($4M) Address ($3M)
W-2 Form	Name Address Date of Birth Social Security Number Phone Number	Address Email Address Driver's License Information	Social Security Number ($22M) Photograph of a Person ($15M) Address ($4M) Date of Birth ($3M) Phone Number ($2M)
Photo of a Person	Bank Account Information Name Address Date of Birth Social Security Number	Phone Number Patient Database Customer Database	Social Security Number ($21M) Email Address ($15M) Address ($5M) Date of Birth ($3M) Phone Number ($1M)

space. In this paper, we focused on three questions: 1) given the evidence of the breach of a subset of identity attributes, what is the impact on the exposure risk of other attributes, 2) given that an attribute has been exposed, what was the most likely source of the exposure, and 3) what is the total cost, including secondary costs, of exposure of a node, and based on this cost, what are the hot-spots in the Identity Ecosystem, i.e., nodes that are both vulnerable to an attack, and carry large liability in case of an exposure.

In this paper, we also focused on the analytics that the Identity Ecosystem provides when using the real world data from ITAP. The insights that the Identity Ecosystem provides are significant, valuable, and sometimes very nonintuitive. Just to mention a few results extracted from the combination of ITAP and Ecosystem, we summarize the following takeaways from the paper.

(1) Digital PII such as Password, Login Credentials, and Employee Login Credentials are widely used in identity theft and fraud.

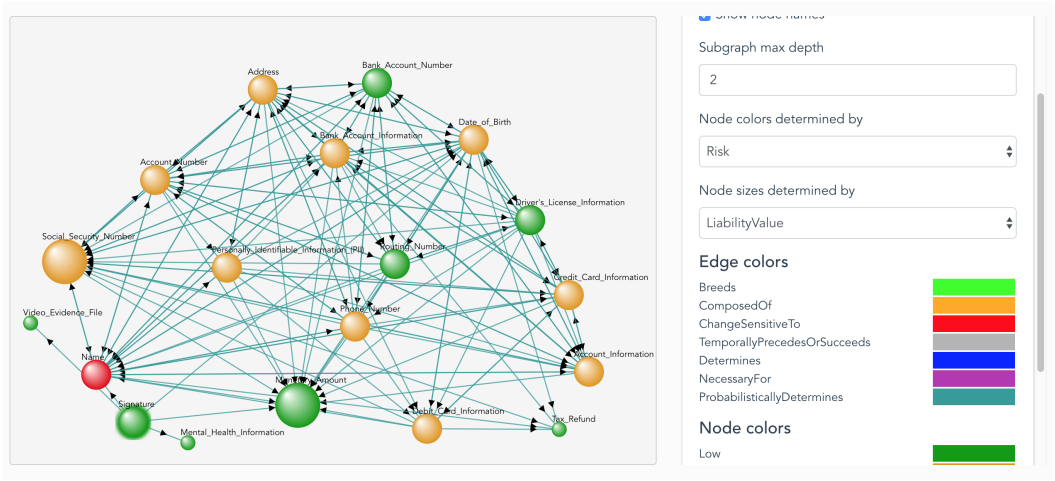

Fig. 4. The Rooted Tree of Depth 2 Rooted at the Signature Attribute.

(2) Digital PII are also targeted in highest profile financial identity theft and fraud cases and are in the top 5% most valuable PII, while, surprisingly, banking information and medical information are not in the top 5%.
(3) In the event of the breach of Credit Card Number, all of the reported likely sources of exposure are either directly related to credit cards, or are digital. Similarly for the breach of Birth Certificate and W-2 Form Information, likely sources of origin are either Driver's Lisence Information or digital login credentials (e.g., Email Address and Employee Login Credentials).
(4) Social Security Number consistently ranks high with high liability in the hot-spots.

Ecosystem provides an endless opportunity to answer similar questions and extract priceless insight, beyond the questions we asked in this paper, to empower users and organizations to combat identity theft and fraud.

REFERENCES

[1] [n. d.]. D3. Retrieved January 16, 2019 from https://d3js.org
[2] Christopher M Bishop. 2006. Pattern recognition. *Machine Learning* 128 (2006).
[3] Erika Harrell, Bureau of Justice Statistics, US Dept of Justice, and Office of Justice Programs. 2019. Victims of Identity Theft, 2016. (2019). https://www.bjs.gov/content/pub/pdf/vit16.pdf
[4] CHICAGO TRIBUNE Joseph Menn. 2007. In identity theft, it's not just about who you know. Retrieved Feb. 15, 2019 from https://www.chicagotribune.com/news/ct-xpm-2007-11-18-0711170017-story.html
[5] Razieh Nokhbeh Zaeem, Monisha Manoharan, and K Suzanne Barber. 2016. Risk Kit: Highlighting Vulnerable Identity Assets for Specific Age Groups. In *European Intelligence and Security Informatics Conference (EISIC)*. IEEE, 32–38.
[6] Federal Trade Commission et al. 2014. Consumer sentinel network data book. (2014).
[7] Yongpeng Yang. 2014. *Mining of identity theft stories to model and assess identity threat behaviors.* Master's thesis. The University of Texas at Austin.
[8] Yongpeng Yang, Manmohan Manoharan, and K Suzanne Barber. 2014. Modelling and Analysis of Identity Threat Behaviors through Text Mining of Identity Theft Stories. In *IEEE Joint Intelligence and Security Informatics Conference (JISIC)*. 184–191.
[9] Jim Zaiss, Razieh Nokhbeh Zaeem, and K Suzanne Barber. [n. d.]. Identity Threat Assessment and Prediction. *Journal of Consumer Affairs* ([n. d.]). https://doi.org/10.1111/joca.12191

Received November 2018; revised November 2018; accepted November 2018

WWW.IDENTITY.UTEXAS.EDU