US-Centric vs. International Personally Identifiable Information: A Comparison Using the UTCID Identity Ecosystem

Rima Rana
Razieh Nokhbeh Zaeem
K. Suzanne Barber

April 2019
UTCID Report #19-01
US-Centric vs. International Personally Identifiable Information: A Comparison Using the UT CID Identity Ecosystem

Rima Rana  
Center for Identity  
The University of Texas at Austin  
Austin, USA  
rima.rana@utexas.edu

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

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

Abstract—Personally Identifiable Information (PII) refers to any information that can be used to trace or identify an individual. A Javelin Strategy and Research Report stated that PII misuse and fraud hits record high with 15.4 million US victims in 2016, about 16% more than the previous year. A comprehensive analysis of PII attributes and their relationships is necessary to protect users from identity theft.

However, identity theft and fraud are not just a US problem. According to a new report from Risk Based Security, in 2016, there were 4,149 confirmed breaches exposing more than 4.2 billion records globally. That is approximately 3.2 billion more records than were exposed in 2013, the previous all-time high.

In this paper, we extend the mathematical representation and implementation model of the UT CID Identity Ecosystem representing PII attributes and relationships to incorporate international PII. Previously, the UT CID Identity Ecosystem model has been primarily populated using data about US theft and fraud cases to include PII attributes used to transact crime as well as accidental exposure of PII attributes. Statistics are also calculated and associated with respective PII attributes such as the frequency of exposure occurrences for respective PII attributes, monetization value of PII (i.e. financial consequences of the crime), and strength of relationships between PII attributes. This research describes how the content of the UT CID Identity Ecosystem and resulting analysis change when PII attributes from international identity theft and fraud cases are incorporated. Not only are the PII attributes different in an international UT CID Identity Ecosystem, the relationships between PII attributes change, the monetization value of PII attributes change, and the risk of exposure change when worldwide identity theft and fraud cases are considered.

Keywords—Personally Identifiable Information, internationalization, identity theft and fraud, UT CID Identity Ecosystem

I. INTRODUCTION

Identity theft is the deliberate or accidental exposure of PII and identity fraud involves the use of a person’s identity, represented by Personally Identifiable Information (PII), to incur financial loss, reputational damage or emotional distress. A person’s PII can encompass data describing both physical and digital attributes that serve to identify the person with the growing intermingling of online and offline attributes. Examples of online attributes are one’s social media accounts, online shopping patterns, passwords, and email accounts. Offline attributes are those related to the physical world such as bank accounts, credit and debit cards, Social Security Number, and one’s physical characteristics. Of course, the physical worlds and digital worlds are connected. A person may hold a credit card but all the information associated with the card and the person are represented and managed in a digital form. Other examples include biometrics that represents physical characteristics of a person (e.g. fingerprints, face, gait) or information tracking one’s physical location that is all managed in a digital form.

The UT CID Identity Ecosystem is a tool developed at the Center for Identity at the University of Texas at Austin that can be used to estimate the value of PII attributes, determine the connectedness and dependencies between PII and predict the risk of losing PII and the liability associated with the fraudulent use of respective PII attributes. The UT CID Identity Ecosystem has every PII attribute modeled as a graph node and PII attribute relationships modeled as graph edges. The “probabilistically determines” relationship between two PII attribute nodes is very informative in this research analysis since it represents the probability that PII attribute A can be determined if PII attribute B is known (exposed). The model uses a Bayesian Belief Network (with cycles allowed) and utilizes Gibb’s sampling to approximate the posterior probabilities of exposure for PII attributes after a breach. When a set of PII attributes is breached, the UT CID Identity Ecosystem can calculate the risk of exposure for PII attributes or identify the hotspots after the breach. For example, the UT CID Identity Ecosystem can calculate the change in the probability of exposure after a Social Security Number (SSN) has been compromised by answering a query titled “Infer probability of breach based on evidence”.

The previous Identity Ecosystem was limited to a model that utilizes PII stories based in the USA. However, with the globally mobile citizens, the identity of an individual affects many people across different geographical locations. A person who travels to Australia from the USA may have multiple addresses. His/her digital device like a mobile phone or laptop, storing personal information or credit card details, can be
exposed in either the USA or Australia. A security incident that happens to the mobile device may expose personal information of family members or related organizations worldwide and result in financial losses across countries. By getting international PII, we obtain a comprehensive knowledge of PII in today’s connected world.

In the following section, we briefly explain the Ecosystem model. Then we discuss our data resource, namely ITAP. and get in depth for US centric PII and International PII. Finally, we present the conclusion and propose future work.

II. ECOSYSTEM MODEL

The Center for Identity at the University of Texas at Austin has constructed a graph-based model whereby the nodes represent individual’s attributes and the edges represents the relationships between these nodes. It provides a statistical framework for understanding the value, risk and mutual relationships of personally identifiable information attributes. Name, age, Social Security Number are examples of such attributes of a person’s identity. Attributes like Driver’s license are unique to a person whereas attributes like country of origin can be common to many people. PII attributes in the form of nodes are related to each other in different ways: breeds, composed of, changes sensitive to, precedes, determines, necessary for and lastly, and probabilistically determines. Probabilistically determines is the relationship used in Identity Ecosystem for this paper to calculate financial consequences for sample scenarios. For example, an individual in the US generally gives his social security number for employment purposes, therefore his/her social security number probabilistically determines his/her employment records. The relationship between two attributes a and b is shown with a directed edge from a to b.

The UT CID Identity Ecosystem uses a Network Model to simulate the relationships among PII for individuals. These models can efficiently perform even in scenarios of missing or incomplete dataset. The graph is visualized as a 3D network graph, which can be rotated, displaced, or moved as and when required.

![Fig 1: Infer the probability of breach based on Social Security Card and Social Security Number as evidence.](image1)

![Fig 2: Risk and Value in Ecosystem with Probabilistically Determines relationship edges](image2)

The model can accurately identify probable nodes of a breach with a probabilistic distribution of breach for each node in the graph. The hot-spots for a possible breach are identified based on these probabilities.

1. Cause of breach: In scenarios where a breach is reported (i.e., an individual’s PII is compromised) the model can identify the most probable origin of breach. A query can be run by selecting the identity information of the individual and identifying the probability of breach at each node.

2. Cost/Liability: In Identity Management, an expense/cost is expenditure on protecting individual’s identity information. The model which generated probability of breach at each node can be utilized to identify a group of hot-spots, or nodes with higher probabilities of breach. These nodes should be protected/secured more than others.

3. Effect of Exposure: The model can also be utilized to compute changes in probability of breach at other nodes, when data at a node is compromised. Bayesian inference model is utilized to compute this change. The expectation of loss of the nodes compromised can also be computed using this model. Moreover, multiple nodes can be used as attributes for computation.

III. DATA SOURCES

The UT CID Identity Ecosystem is populated by modeling stories of real world theft and fraud cases in the UT CID Identity Threat Assessment and Prediction (ITAP) project. The UT CID ITAP is a risk assessment tool, which collects case data from sources like law enforcement and the news media and has covered almost 6,000 identity theft and fraud scenarios. A team of modelers at the Center for Identity analyzes identity theft news and stories daily to model the value of identity attributes and their risk of exposure. Specifically, the ITAP model describes the business process (inputs, process steps, outputs, consequences, and victims impacted) by which PII is deliberately stolen, accidentally exposed, and fraudulently used. Earlier, the UT CID ITAP focused on stories in the US, however, now with the inclusion of international PII attributes, it gets global with around 12.3%
(and increasing) international cases. The UT CID ITAP stories coverage is extensive and ranges from Australia, New Zealand, and the UK to China and India with also some representation from South Africa, South Korea, Singapore, Thailand, Middle East and even cybercrimes that are independent of the geographical location.

The UT CID ITAP analyzes fraud cases based on market sector, demographics of victims and now victims from different countries along with the US to compute a model of identity to provide a unique global calculation of risk of exposure and intrinsic loss value.

A. US Centric PII

The previous implementation of the UT CID Identity Ecosystem used ITAP identity theft and fraud stories based in the United States, enabling sophisticated inference about people, devices and organizations in the identity space. Financial consequences of exposure of the PII attributes in US case is calculated and used to compare with the international PII.

Table I: Classification of nodes for US Scenarios

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyber/ Online: Email account, IP Address, User Credentials</td>
<td>Employer: IRS Documents, W2 Form, Job Title</td>
<td>Family: Spouse PII, Relative’s PII</td>
</tr>
</tbody>
</table>

B. International PII

Identity breach is now a widespread problem across the globe with increasing degree of sophistication in hacking involved. The UT CID Identity Ecosystem gets input from ITAP considering international hacks, attacks and breach stories. This helps in evaluating the risk and effect on other PII for an individual or related people, devices or organizations. The financial consequences involved in exposure of international PII are multiplicatively enhanced with an organization having customer base across continents risking identities of many people.

To understand the Personally Identifiable Information in International context, identity thefts in different countries are considered but also cyber world attacks impacting individuals from more than one country. Recent data breach from major internet companies, phishing scams, hacking, server breaches, malware attacks and many other different types of cyberattacks are compromising the online PII of an individual which violates privacy in cyber as well as physical world.

Though some of the identity attributes are common between US centric PII and international PII, the impact of the exposure and risk varies numerically and even the relationships, i.e. the edges, change. For example, PII node of Social Security Number can be used to identify US PII uniquely but can be associated with a person belonging to another nationality in the international PII and multiplicatively increase the monetary consequences given the SSN PII is breached as compared to the same pretext in US PII scenario.

International PII for organizations will have differences in impact of exposure of any PII attribute node depending on the deployment and data storage solutions. The risk and financial loss involved in the international PII node exposure for an on-premise solution is observed to be less as compared to an organization having cloud solution.

C. Comparative Analysis

Previously, the UT CID Identity Ecosystem had focused on US theft and fraud stories. This research describes how the content of the UT CID Identity Ecosystem and resulting analysis change when PII attributes from international identity theft and fraud cases are incorporated. To incorporate internationalization into the UT CID Identity Ecosystem, the international PII attributes are introduced from international theft and fraud stories. Consequently, the values of all PII attributes, US and international, change in terms of monetary loss details for global breaches, and so does the existing relationships between PII attributes and the risk of PII exposure resulting from a data breach. The PII attributes are divided into three categories based on scope, value and risk for this global examination of PII: US Specific, Country Specific and Globally Prevalent. Distinguishing country specific and globally prevalent PII attributes, as well as the relationship between PII attributes in the international context, improves our understanding and fine-tunes the calculation of probabilities. This research aims to enhance our understanding of the scope, value, and risk levels of PII attributes in the international context, thereby, improving our understanding of how best to protect these PII attributes from theft and fraud.

The categorization of PII attributes helps in understanding how change of geographical location affects certain PII attributes and does not affect others. A person’s online PII attributes like email accounts and social media accounts remain globally the same. On the other hand, PII attributes pertaining to tax or medical information are country specific. Such country specific attributes tend to change, based on the geographical location.

<table>
<thead>
<tr>
<th>Countries covered for International Scenarios</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afghanistan</td>
</tr>
<tr>
<td>Balkans</td>
</tr>
<tr>
<td>Brazil</td>
</tr>
<tr>
<td>Colombia</td>
</tr>
<tr>
<td>European Countries (Other)</td>
</tr>
<tr>
<td>India</td>
</tr>
<tr>
<td>Italy</td>
</tr>
<tr>
<td>Middle East</td>
</tr>
<tr>
<td>Pakistan</td>
</tr>
<tr>
<td>Russia</td>
</tr>
<tr>
<td>South Africa</td>
</tr>
<tr>
<td>Switzerland</td>
</tr>
<tr>
<td>Turkey</td>
</tr>
<tr>
<td>Vietnam</td>
</tr>
</tbody>
</table>

Copyright 2018 The University of Texas. Confidential and Proprietary, All Rights Reserved.
in the value of PII attribute as well as the effect on the edges with other PII attributes and hence the risk of exposure.

Every PII Attribute in both the ITAP Scenarios have the probability of getting exposed from the "probabilistically determines" edge relationship which is calculated as the risk of exposure. Financial consequences relate to the dollar value associated with the theft of that PII according to the ITAP stories collected. Based on the identity theft covered in the US ITAP scenario stories and international ITAP stories, analysis can be drawn about the risk of exposure and financial losses associated with PII attribute depending on the category to which it belongs.

Observing some of the PII attributes of each category (US specific, country specific and global), a trend can be seen for risk of exposure and financial consequences between US and International ITAP. Attributes belonging to the Global PII like Date of Birth and IP Address, increase both in risk of exposure and financial consequences as the ITAP theft stories go from US based to international. As seen in Table IV, PII attributes like Zip code and Visa details, which are country specific, have a reduction in the risk of exposure; however, they witness an increase in the financial loss. Lastly, while PII attributes relative to US such as W-2 form or National Identity Number have an increase in risk of exposure, they show a slight decrease in monetary losses.

Table IV: Risk of Exposure and Financial Consequences for US and International ITAP for some PII attributes.

<table>
<thead>
<tr>
<th>PII Attribute Type</th>
<th>Risk of Exposure (US ITAP)</th>
<th>Risk of Exposure (International ITAP)</th>
<th>Financial Consequences in USD (US ITAP)</th>
<th>Financial Consequences in USD (International ITAP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP Address/Global</td>
<td>0.001456</td>
<td>0.000534</td>
<td>3822967</td>
<td>7223816</td>
</tr>
<tr>
<td>Date of Birth/Global</td>
<td>0.233261</td>
<td>0.000178</td>
<td>6275159</td>
<td>10835724</td>
</tr>
<tr>
<td>Zip Code/Country specific</td>
<td>0.003275</td>
<td>0.001246</td>
<td>116453</td>
<td>4728544</td>
</tr>
<tr>
<td>Visa details/Country specific</td>
<td>0.000364</td>
<td>0.001246</td>
<td>15291869</td>
<td>6328318</td>
</tr>
</tbody>
</table>

Looking at the nodes used for international and US based ITAP data in Table V, we observe the number of PII nodes have almost doubled. The number of edges, probabilistically determines relationships have increased to almost thrice. The percentage of connected nodes remains constant at 78% however with the increase in number of edges, the financial impact of exposing one PII attribute increases multiplicatively in international scenarios.

Table V: ITAP Data used for Ecosystem

<table>
<thead>
<tr>
<th></th>
<th>International ITAP</th>
<th>US ITAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Nodes</td>
<td>542</td>
<td>296</td>
</tr>
<tr>
<td>Number of Edges</td>
<td>4742</td>
<td>1758</td>
</tr>
<tr>
<td>Percentage of Connected nodes/Total nodes</td>
<td>78.04%</td>
<td>78.37%</td>
</tr>
</tbody>
</table>

D. Impact of GDPR on PII

The General Data Protection Regulation aims to give control of personal data to the citizens of the European Union. After its implementation on May 25 2018, it provides rules and procedures for the processing of the Personally Identifiable Information of individuals inside the EU and applies to all the organizations doing business in European Economic Area, irrespective of location. Hence, GDPR can reduce the value of monetary loss propagated by the exposure of international PII.

IV. CONCLUSION AND FUTURE WORK

This work seeks to improve our understanding of PII attributes in the international context. It hence enhances our ability to protect these PII attributes from theft and fraud. In this paper a novel concept of international PII was introduced which offers an insight into how personally identifiable information is utilized in different countries. We are interested in the differences between PII in the US context versus that of the international one. Previously, the UT CID Ecosystem could answer queries related to people, devices and organizations for the USA. By combining and analyzing different identity theft stories across the globe, this paper derives more comprehensive and holistic results from the UT CID Identity Ecosystem. The potential coverage from ITAP needs expansion to predict more accurate PII for the international context since the UT CID Ecosystem envisions to identify and authenticate people, devices and organizations internationally. In the future, we plan to incorporate international ITAP stories more extensively and extend the queries that the UT CID Ecosystem answers.

ACKNOWLEDGMENT

The authors would like to thank James Zaiss for his help on the UT CID ITAP work.
REFERENCES


