



All Theses and Dissertations

---

2017-12-01

# An Analysis of the Size and Impact of Digital Footprints

Whitney Nielsen Maxwell  
*Brigham Young University*

Follow this and additional works at: <https://scholarsarchive.byu.edu/etd>



Part of the [Systems Engineering Commons](#)

---

## BYU ScholarsArchive Citation

Maxwell, Whitney Nielsen, "An Analysis of the Size and Impact of Digital Footprints" (2017). *All Theses and Dissertations*. 6593.  
<https://scholarsarchive.byu.edu/etd/6593>

This Thesis is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in All Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact [scholarsarchive@byu.edu](mailto:scholarsarchive@byu.edu), [ellen\\_amatangelo@byu.edu](mailto:ellen_amatangelo@byu.edu).

An Analysis of the Size and Impact of Digital Footprints

Whitney Nielsen Maxwell

A thesis submitted to the faculty of  
Brigham Young University  
in partial fulfillment of the requirements for the degree of  
Master of Science

Chia-Chi Teng, Chair  
Dale C. Rowe  
Derek L. Hansen

School of Technology  
Brigham Young University

Copyright © 2017 Whitney Nielsen Maxwell

All Rights Reserved

## ABSTRACT

### An Analysis of the Size and Impact of Digital Footprints

Whitney Nielsen Maxwell  
School of Technology, BYU  
Master of Science

Personal information available online is known as a digital footprint. While many have a digital footprint, few if any, know what it encapsulates or how to control it. Technology and personal information are becoming more intertwined as technology becomes more integrated with everyday activities. Personal information can be defined as details that apply to a person such as race or shopping habits. Shopping habits are considered personal information by many corporations who spend money to track, or even predict purchases of individuals, whereas more traditional forms of personal information are details like gender, birthdate, and home town. With a wide breadth of personal information available, not all of it is equally valuable or personally unique. This project is dedicated to determining the content and size of a digital footprint, and assessing its impact for an individual by defining the discoverability of that content.

Keywords: information security, digital footprint, personal information, information privacy

## ACKNOWLEDGEMENTS

I would like to thank my graduate committee and all of the BYU IT faculty for their support to accomplish this degree remotely, and for my employer Microsoft, who provided funding for my tuition. I would also like to thank my family, specifically my parents and my sisters, who were a never-ending source of faith and encouragement. Lastly, I would like to thank my husband Tim, who always believed in me and became my biggest supporter. I couldn't have done it without all of you.

## TABLE OF CONTENTS

<b>LIST OF TABLES .....</b>	<b>vi</b>
<b>LIST OF FIGURES .....</b>	<b>vii</b>
<b>1 Introduction .....</b>	<b>1</b>
1.1 Nature of Problem .....	1
1.2 Purpose of Research .....	2
1.3 Project Approach.....	3
1.4 Research Questions .....	3
1.5 Definitions.....	4
<b>2 Literature Review .....</b>	<b>5</b>
2.1 Categories and Classification of Personal Information.....	5
2.2 Price of Personal Information .....	6
2.3 Personal Information Leakage Online.....	8
<b>3 Methodology.....</b>	<b>9</b>
3.1 Project Design .....	9
3.2 Institutional Review Board.....	9
3.3 (R1) Data Collection .....	11
3.3.1 Layers.....	11
3.3.2 Weights .....	11
3.3.3 Data Organization.....	12

3.3.4	Research Steps .....	17
3.4	(R2) Data Analysis: The Personal Information Vulnerability Assessment Score .....	19
<b>4</b>	<b>Digital Footprint Collection and Analysis .....</b>	<b>25</b>
4.1	Research Results .....	25
4.2	Analysis of Identifiers .....	27
4.3	Analysis of Categories .....	33
4.4	Analysis of the Research Participants: PIVA Score.....	36
<b>5</b>	<b>Conclusions and Future Work .....</b>	<b>39</b>
5.1	Project Contributions.....	40
5.2	Future Research.....	41
5.3	Potential Applications .....	42
	<b>References .....</b>	<b>44</b>
<b>Appendix A.</b>	<b>Research Results.....</b>	<b>46</b>
<b>Appendix B.</b>	<b>IRB Documents .....</b>	<b>59</b>

## LIST OF TABLES

Table 3-1: List of Layers and Categories.....	12
Table 3-2: List of Resources Used for Research .....	15
Table 4-1: Success Rates for Layers Given Email and Phone Number.....	28
Table 4-2: PIVA Scores for Research Participants.....	36

## LIST OF FIGURES

Figure 3-1: Demographic of Research Participants .....	14
Figure 3-2: Data Collection Format.....	16
Figure 3-3: Example of Including Source in Data Collection .....	17
Figure 3-4: Venn Diagram of Identifiers .....	18
Figure 3-5: Research Phases .....	18
Figure 3-6: FICO and PIVA Weight Distribution .....	21
Figure 3-7: Example Set of Results .....	22
Figure 4-1:RP_G Results .....	26
Figure 4-2: RP_G Results Continued .....	27
Figure 4-3: Research Participant RP_A Results Using Only Phone Number .....	29
Figure 4-4: Research Participant RP_H Results Using Only Email Address.....	30
Figure 4-5: Categories Discovered from Phone and Email Separately .....	31
Figure 4-6: Mapping of Categories to Email Address and Phone Number .....	32
Figure 4-7: Weighted Mapping of Categories to Phone and Email.....	33
Figure 4-8: Probability by Category .....	35
Figure 4-10: Normal Distribution of PIVA Scores.....	37
Figure 4-11: PIVA Score for RP_A.....	38



# 1 INTRODUCTION

## 1.1 Nature of Problem

As users have adopted technology over the ages, technology has evolved to become more personable and “smart.” With that personalization, users have begun to trust technology with more and more of their own personal information without realizing that they are creating a bigger and deeper digital footprint. Increasingly, people still want to use technology, but they don’t want to have their personal information exposed as they use it.

If knowledge is power, then information is the new currency. Privacy and anonymity are no longer concerns that are limited to criminals or the rich and famous; they affect everyone. Events like Edward Snowden’s information leak about NSA surveillance on US residents spark questions regarding what information (if any) is kept private (Maxwell, 2016). Currently, there is a gap between users’ trust in invasive technology and their understanding of the amount of personal data those technologies reveal publicly (Lee, 2015). There is resistance against having a “Big Brother,” yet few realize that more than one already exists (Schneier, 2015). While awareness of privacy has increased in recent years, little is understood about how to monitor, control, and redact information that is already online.

Technology and personal information seem to go hand in hand. It is just as common for a person to register their email address on a pizza website as on a banking website. Since so many facets of technology access various pieces of personally identifiable information, it is

almost impossible to keep track of exactly where and how much personal information is publicly accessible. Furthermore, once a piece of information becomes public, it can be difficult to redact. With an ever-increasing adoption of technology, it is important to become aware of what personal information is revealed.

## **1.2 Purpose of Research**

This research aims not only to educate and inform users on the breadth and depth of their digital footprint, but to reform how they interact with online services to better protect their personal information. Personal information should be just that, personal. This research will address two different problems: first, assessing the amount of personal information currently exposed, and second, assessing the impact of the exposure.

Understanding what information is currently exposed is the primary inspiration for this research. Even individuals who have taken a vested interest to protect their personal information, find it's not easy to keep track of it all. In addition to discoverability of personal information, the relationships of various digital information can also be meaningful. Instead of searching for all the places that disclose information, why not try to discover what a single piece of information discloses? In effect, this research not only reveals insight into the amount of personal information exposed, but sheds light onto which pieces of information reveal the most.

To assess the impact of personal information, there needs to be a ranking of sensitivity for different areas of personal information. This research project assigns a value to information based on its level of uniqueness to an individual. Finally, this project assesses the vulnerability of personal information, via research on its discoverability therefore increasing awareness.

### **1.3 Project Approach**

The goal of this project is to research digital footprints of a sampling of the population that uses online services and then analyze that data for trends and patterns on how different layers connect to make up a digital footprint.

Naturally for some users, uncovering information would be easier if they have public profiles on multiple social networking sites. Other users, will find it more challenging to uncover the same information. As a result, different profiles of the common digital footprint may emerge for various demographics such as millennial, youth, and middle aged.

### **1.4 Research Questions**

The following research questions will be addressed:

- (R1) How much information can be uncovered about a person given one or two details?
- (R2) How can a digital footprint be captured and measured in size and impact using various metrics?

The proposed research will also uncover personal information of people in various demographics. This will help analyze and develop a metric of measuring digital footprints.

- Hypothesis (H1): Details that are generally unique to an individual yet common (e.g. phone number and email address) reveal personal information. Because these details are used in association with areas like work, school, home, and hobbies, it's expected they would lead to information that will reveal a great deal about each of those areas.

## 1.5 Definitions

- Digital Footprint - Information about a particular person that exists on the Internet as a result of their online activity.
- Personal Information - Recorded information about an identifiable individual.
- Vulnerability - A weakness or area that is exposed or at risk.
- Assessment - Evaluation or estimation of the nature, quality, or ability of someone or something.
- Data Mining - Practice of examining large databases in order to generate new information.

## **2 LITERATURE REVIEW**

Numerous studies and academic articles exist on information privacy, digital footprints, information leakage via social networking, and price of information. This chapter will analyze current studies that are relevant to this research.

### **2.1 Categories and Classification of Personal Information**

The term “personal information” can be vague since it can vary from culture to culture or even person to person (Otsuki, 2013). In order to be a bit more transparent for research, it is best to structure and classify the various aspects of what defines “personal information.” Literature suggests that personal information can be divided into three categories or layers: peripheral, intermediate, and core (Shibchurn, 2014). The peripheral layer (also called the identifier layer) is composed of biographic data such as name, age, and gender. The intermediate layer (also called the demographic layer) is composed of affiliations such as religion, political views, and hometown. The core layer (also known as the favorites and tastes layer) is composed of emotions and values such as favorite music/books, sports, and activities.

Clearly, some information is inherently more sensitive than others. According to a study at Beijing University, “Appropriately classifying information is the basis for preparing formation control strategy and for sharing information” (Shi, 2007). This classifies personal information based on its value and sensitivity. Value is determined through elements such as reliability,

increment, timeliness, availability, cost of creation, cost of service, cost of usage and cost of opportunity. Sensitivity is determined through elements such as tolerance of exposure and spread, cost of regeneration, and time of regeneration.

The problem is that the classification and separation of information is based on theories not research. There is no attempt to actually discover information and verify sensitivity and value. This research will use these methods of classifying and categorizing personal information on collected data to compare sensitivity and discoverability.

## **2.2 Price of Personal Information**

As mentioned previously, users are becoming more aware that their information is sold for marketing and profit to third parties and has led a number of parties to voice privacy concerns. “The U.S. Federal Trade Commission has noted that data collection can be invisible, privacy notices may be difficult to understand, consumer profiles are sometimes very detailed, and that there is a ‘risk that data collected for behavioral advertising – including sensitive data regarding health, finances, or children – could fall into the wrong hands or be used for unanticipated purposes’” (Ur, 2012). In an effort to gauge user’s interest on shouldering cost to improve information privacy, there has been some research to measure just how much users would be willing to pay. According to a study in New Zealand, “New Zealanders were willing to pay NZD 55.40 (USD 28.25) for property rights to protect privacy. [It’s] estimated that individuals in the US sample valued privacy at USD 30.49 to 44.62 while those in the Singapore sample valued it at USD 10.45 to 26.93” (Rose, 2005). Similarly, other studies have found that users value their browsing history for about 8\$ (Kokolakis, 2017).

A study of students from the University of Salerno found, “when asked to choose between search quality and search privacy, 61% expressed their preference for privacy” (Malandrino, 2013). Yet, preference does not always mean action. In fact, observations of behavior in the marketplace suggest that people who express concern over their personal information can be less selective in its protection and is known as the “privacy paradox” (Norberg, 2007 & Dienlin, 2015). The discrepancy between preference and action can partially be tied to gender and age. Women tend to be more concerned about privacy, yet men are more likely to adopt protective behaviors (Bartel, 1999) and younger users are more likely to protect their privacy than older ones (Kokolakis, 2017).

A recent study argued that online privacy behaviors are not paradoxical, but are based on distinct privacy attitudes (Dienlin, 2015). These attitudes could mean that users are motivated to protect their information, but they lack understanding of how and what to protect. Users who employ privacy-enhancing behaviors also have more technical skills related to privacy. In short, if people want to establish self-protecting behaviors they need a sufficient amount of technical skills related to privacy (Dienlin, 2016). Studies have shown, “that users are willing to adopt simple and easily applicable strategies, but not complicated tools which require advanced technical skills or consume a lot of time” (Matzner, 2017). This project will produce results that aim to educate non technical users on their current state of information vulnerability and to understand which areas are at risk. Data privacy shouldn’t be limited to those who can understand protection technologies.

### 2.3 Personal Information Leakage Online

Users in this digital age are all living two lives: offline and online. Paper trails and digital trails coexist making it difficult to segregate where information originates and resides.

Correlating digital information with an offline identity presents a risk of identity thefts, profile cloning, compromised accounts, spam and phishing, online profiling, and online stalking to name a few (Malhotra, 2012). The exponential growth of “smart” technologies has led to a socio-technical environment in which tracking, data mining, and profiling have emerged (Matzner, 2017). The source of this digital information does originate from the user, however, it’s often without their knowledge. When a user visits a website, 56% of the sites directly leak private information (Krishnamurthy, 2011). Leaked data can include medical, financial, name, email address, family, and other sensitive information (Malandrino, 2013).

Leaked information may or may not be damaging on its own, but it’s possible for a malicious attacker to use that data across several sites and correlate it to a single user. According to a study on correlating pseudonyms for social networking sites, “an attacker will find over 60% of the user’s social networking profiles in the best-case (and more than 35% of the profiles in the worst-case)” (Irani, 2009).

These studies concentrate on tool development (such as RequestPolicy, Ghostery, NoScript, and Adblock Plus) to reduce the amount of leakage. Limiting information exposure is key in controlling personal information. Awareness of data collection is crucial in order for users to implement information protection practices because, “it’s possible that users either are not aware of privacy management strategies or find them too burdensome to employ” (Wisniewski, 2014). This research builds on the idea of exposing where and how information is leaked to educate users on how to control it.



### **3 METHODOLOGY**

This chapter outlines the plans for the study and describes how answers for each purposed research question will be found.

#### **3.1 Project Design**

The purpose of this research is to provide a template of a digital footprint that could be applied to users outside of the research study. It is also designed to research how to find personal information from a sampling of users from various demographics, and then analyze the results to look for patterns or generalizations that could be applied to a broader population. The analysis will provide users an insight into the current status of their personal information vulnerabilities and where they can better protect it.

#### **3.2 Institutional Review Board**

The use of human subjects requires the approval of the Institutional Review Board (IRB). According to the IRB, "human subjects are defined as "living individual(s) about whom an investigator conducting research obtains (1) data through intervention or interaction with the individual; or (2) identifiable private information." (Activities, 2017) Since this research project intended to research 5-10 individuals, it required IRB approval.

As part of the application process, each participating faculty member and researcher submitted a brief biography of themselves and completed the Collaborative Institutional Training

Initiative (CITI) certification course, scoring at least 80% in all sections fulfilling the ethics training requirement.

The application itself went through multiple review sessions with the IRB before gaining approval. The approval permitted the study to determine the quantity and quality of information that can be uncovered through a combination of three common identifiers: email address, phone number, and name. The name identifier primarily benefits the researcher, so that there is some assurance that information is being gathered about the right person.

The research employs various Internet searches and research techniques to use those identifiers to uncover additional information online. Personal information available online is what is known as a "digital footprint." Though a digital footprint technically encompasses all personal information available online, this research will only search for predetermined categories that vary in sensitivity. The categories are: additional phone numbers (cell/home/work), home address, photo, hometown, current city, religion, employers, graduate school/college, high school, favorite music, favorite books, favorite movies, favorite TV shows, favorite sports/teams/athletes, and hobbies.

The analysis encompasses the ease or difficulty in uncovering these categories as well as the value of each category. Given the sensitivity of this research, all data is encrypted and secured while it is retained for three years as required by the IRB. The approval period spans from March 28, 2017 to January 11, 2018. IRB application and research participation consent form are listed in Appendix B.

### **3.3 (R1) Data Collection**

#### **3.3.1 Layers**

As mentioned in chapter 2.1, personal information can be divided into three main layers: peripheral, intermediate, and core (Shibchurn, 2014). These layers also represent a gradient of how ‘personal’ the information is. The core layer represents information that could apply to a great number of individuals across cultures, demographics, and time. Still, the intermediate layer is applicable to a large number of individuals, but it represents a smaller group than the core layer. The peripheral layer contains information that generally identifies a specific individual. Research categories are divided into these three layers according to their characterizations and each layer is weighted according to its level of information sensitivity. Table 3-1 shows how categories are divided among the three layers.

#### **3.3.2 Weights**

Information privacy can have different meanings among individuals. To one person, public knowledge of their birthday and favorite movie is acceptable, but to someone else, it’s not acceptable. While this research cannot accommodate each individual’s personal values, it still can recognize the fact that not all information is equally personal.

This project will take into account the specificity of each layer in the analysis by assigning a weight to each layer. The peripheral layer will be weighted higher than the intermediate layer, and the intermediate layer will be weighted higher than the core layer. The higher weight means that information tends to be more unique and bound to the individual. There

are always exceptions to the rule, but this research will rely on previously defined generalizations. More details of the weight structure are found in the section 3.4.

### 3.3.3 Data Organization

For this project, the digital footprint will contain a list of six items within each of the three categories. Searching for items in each category allowed broad enough searches to retain flexibility, but scoped the research to a definitive target. In a sense, the research could be classified as a capture-the-flag of sorts for each individual. Additional categories of information discovered along the way were noted with the rest of the participant data to be analyzed. Table 3-1 displays the list of categories divided by layer:

Table 3-1: List of Layers and Categories

Layer	Category
<b>Peripheral Layer</b>	Birthday
	Name
	Email Address
	Cell Phone/Other Phone #
	Home Address
	Photo
<b>Intermediate Layer</b>	Hometown
	Current City
	Religion
	Employers
	Graduate school/college
	High School
<b>Core Layer</b>	Favorite Music
	Favorite Books
	Favorite Movies
	Favorite TV Shows
	Favorite Sports/Teams/Athletes
	Hobbies
<b>Grand Total</b>	<b>18</b>

There are six categories of information within each of the three layers, making a final total of eighteen categories. The categories were chosen because they can be (and often are) disclosed via social networking sites such as Facebook, Twitter, Instagram, Pinterest, etc. The categories can help determine trends across research subjects without degrading the unique quality of the information. The goal of the research is to try and discover as many of those eighteen categories as possible online for each research participant.

Participants for this research were selected through personal referrals. To take part in the research, all research participants provided their name, email address, and phone number. The reasoning behind the choice of these identifiers include:

- Each research participant would most likely have all three identifiers.
- These identifiers would cause a low amount of discomfort for research participants to disclose.
- These identifiers are most often used in online resources as search parameters.
- Identifiers were unique enough to the individual that they actually increased confidence in search result accuracy.

Figure 3-1 shows the demographic of research participants. There were 3 males and 4 females. Three were aged 18-24, one was 25-35, two were 35-45, and one was 65+. No participants were between the ages of 45-65. Since all of the participants were provided as a referral, there was no way to know ahead of time what the final demographic layout would be.

Most of the participants were referred by a single individual. That individual knew both the researcher and the referred research participants. Most of the participants were referred by a single individual. That individual knew both the researcher and the referred research participants.



Figure 3-1: Demographic of Research Participants

During the research, this individual was able to assist the researcher by answering questions about the research participants and verify the accuracy of search results along the way. The final results were verified with each research participant at the conclusion of the project during a brief phone call to ensure accuracy of the results.

It was necessary to create a similar research experience for each research participant, so that if needed, future research studies could replicate the results. To accomplish that, the same resources were used for each person. Table 3-2 represents the final list of sites that were used on all participants:

Table 3-2: List of Resources Used for Research

Online Resource Name
<a href="http://Instantcheckmate.com">http://Instantcheckmate.com</a>
Maltego [ <a href="https://www.paterva.com/">https://www.paterva.com/</a> ]
<a href="https://intelius.com">https://intelius.com</a>
<a href="https://haveibeenpwned.com/">https://haveibeenpwned.com/</a>
<a href="http://www.zabasearch.com/">http://www.zabasearch.com/</a>
<a href="http://premium.whitepages.com">http://premium.whitepages.com</a>
<a href="http://www.usernamecheck.com/">http://www.usernamecheck.com/</a>
<a href="https://namechk.com/">https://namechk.com/</a>
<a href="https://spokeo.com">https://spokeo.com</a>
<a href="https://pipl.com">https://pipl.com</a>
<a href="https://Knowem.com">https://Knowem.com</a>
<a href="http://123peoplesearch.com">http://123peoplesearch.com</a>
<a href="http://truthfinder.com">http://truthfinder.com</a>
<a href="http://Publicrecords.directory">http://Publicrecords.directory</a>
<a href="http://Dobsearch.com">http://Dobsearch.com</a>
<a href="https://www.truepeoplesearch.com">https://www.truepeoplesearch.com</a>
<a href="https://nuwber.com/">https://nuwber.com/</a>

This list of resources was curated from different sources. Several were taken from the book, “How to Disappear.” The issue with any publication, whether books or articles, are the websites that become broken or obsolete. They change at a rapid pace and it makes it difficult to maintain a resource list that’s accurate. For example, at the beginning of the research, there was a site called <http://birthdatabase.com>. It was an incredibly resourceful site; however, the site was taken down before the research finished. Meanwhile, other sites were discovered during the research period that didn’t exist when it began. Whenever a new site was discovered, it was added to the list and used on all participants.

Among websites listed in the resources, is the tool Maltego. Maltego is an open source intelligence and forensics application, which offers data mining and information gathering. Maltego has various third-party modules that use APIs to perform searches on a particular piece of information, such as an email address. When provided an email address, Maltego searches for a user mapped to that email address from various social networking sites such as Facebook, Pinterest, Twitter, and Flickr. It also verifies the validity of the email address and attempt to map it to a phone number. The scope of this research did not allow for the researcher to attempt to ‘friend’ the participant or take any action to attempt to get further access on a social networking site by contacting the research participant in any way. The only method in scope was Open-Source Intelligence Gathering (OSINT).

As categories were found in each resource, they were tracked in the following format in OneNote as seen in Figure 3-2:

<p><b>Peripheral Layer</b> The peripheral layer (also called the identifier layer) is composed of biographic data such as name, age, and gender.</p> <ul style="list-style-type: none"><li><input type="checkbox"/> Birthday</li><li><input type="checkbox"/> Name</li><li><input type="checkbox"/> Email Address</li><li><input type="checkbox"/> Phone#</li><li><input type="checkbox"/> Home Address</li><li><input type="checkbox"/> Photo</li></ul> <p><b>Intermediate Layer</b> The intermediate layer (also called the demographic layer) is composed of affiliations such as religion, political views, and hometown.</p> <ul style="list-style-type: none"><li><input type="checkbox"/> Hometown</li><li><input type="checkbox"/> Current City</li><li><input type="checkbox"/> Religion</li><li><input type="checkbox"/> Employers</li><li><input type="checkbox"/> Graduate school/college</li><li><input type="checkbox"/> High School</li></ul> <p><b>Core Layer</b> The core layer (also known as the favorites and tastes layer) is composed of emotions and values such as favorite music/books, sports, and activities.</p> <ul style="list-style-type: none"><li><input type="checkbox"/> Favorite Music</li><li><input type="checkbox"/> Favorite Books</li><li><input type="checkbox"/> Favorite Movies</li><li><input type="checkbox"/> Favorite TV shows</li><li><input type="checkbox"/> Favorite Sports/Teams/Athletes</li><li><input type="checkbox"/> Hobbies</li></ul>
---

Figure 3-2: Data Collection Format



When a category was found (e.g. birthday), the URL of where it was found was included for further tracking and data analysis as seen in Figure 3-3:

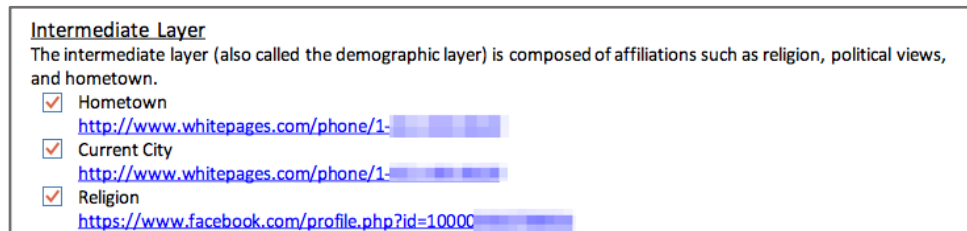


Figure 3-3: Example of Including Source in Data Collection

Data is tracked this way so personal information of research participants are not duplicated to an additional source (which would cause further vulnerability to the research participant), and thereby satisfy requirements set by the IRB. Noting the source of the information disclosure is valuable for the analysis and to the research participant.

### 3.3.4 Research Steps

Initially the research was going to be done by starting with the outer circles of the Venn diagram and working inward as shown in Figure 3-4. The order would have been:

- Name
- Email Address
- Phone Number
- Name and Email Address
- Name and Phone Number
- Email Address and Phone Number
- Name and Phone Number and Email Address

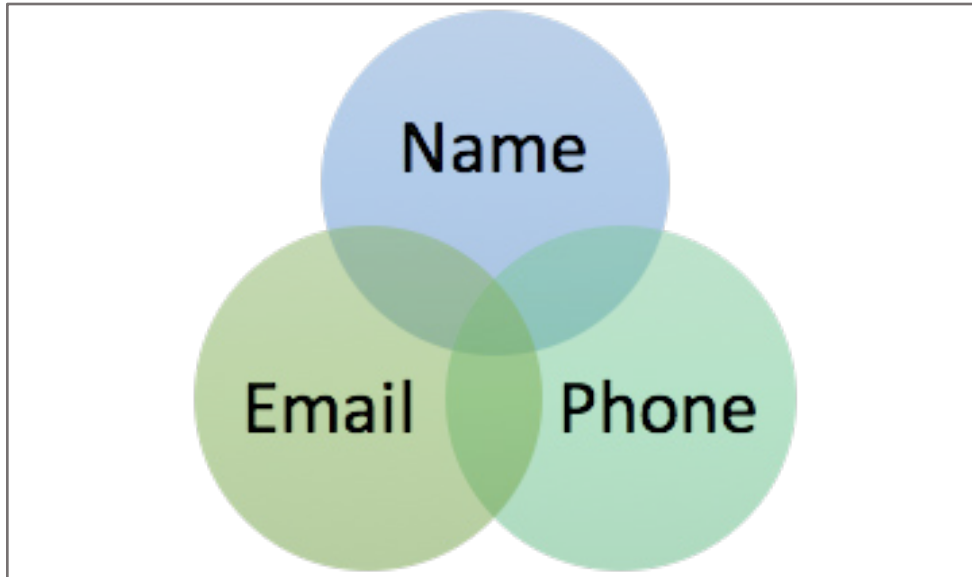


Figure 3-4: Venn Diagram of Identifiers

However, keeping these strict combinations wasn't possible since most sites don't allow for multi-parameter searches and if they do, it's by name and location, not email address or phone number. Instead, the research was done in the following phases as seen in Figure 3-5:

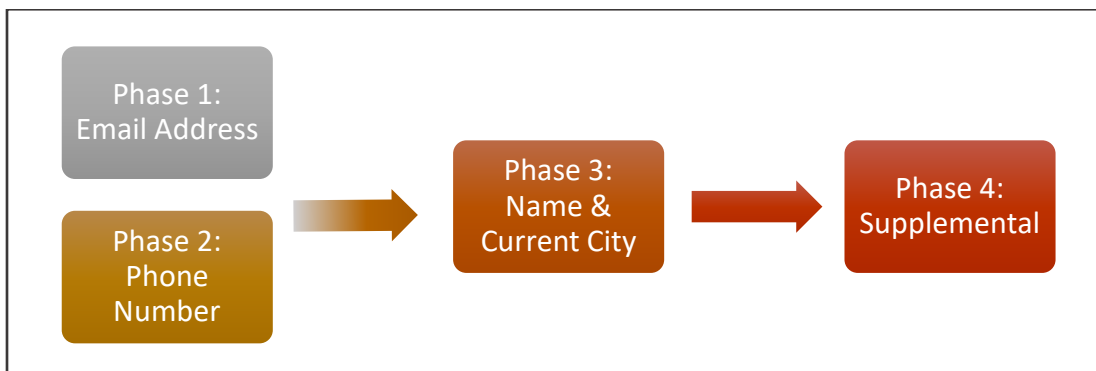


Figure 3-5: Research Phases

The first phase is email. The email address was used across all sites from the resource list that accepted an email address as a search parameter. All discovered categories were ticked off

the list for that research participant in OneNote as shown in Figure 3-3, and the URL was noted below. The same process was followed for phase two using the phone number. Once the first two identifiers had been searched independently, their results were combined for the remaining phases. If found from the previous phases, phase three uses the current city with the name as search parameters. Once all initial identifiers were used across the list of online resources, the research moved on to phase four, supplemental.

The supplemental phase consisted of taking additional research steps to uncover categories that might not have applied to each research participant. This phase is dependent on information previously gathered in phases one and two, and demonstrated information gathering for a research participant often takes a mix of resources and methods. For instance, using discovered photos can retrieve photo metadata such as the date the photo was taken, the camera that took the photo, or even the GPS coordinates. Gmail can resolve an @gmail.com address to a name and photo. Information found about relatives or associates can also uncover missing information. Oftentimes, searching for a relative or associate reveals information about the research participant. The supplemental phase of research was often very successful, however, due to the multifaceted searches, retracing categories back to identifiers was messy.

### **3.4 (R2) Data Analysis: The Personal Information Vulnerability Assessment Score**

The inspiration for the Personal Information Vulnerability Assessment (PIVA) score is the Fair Isaac Corporation score, also known as a FICO score. Its function is to calculate credit scores for individuals. The calculation for the FICO score is comprised of five different weighted areas of a person's credit, and then added to provide a total score. The top FICO score is 850, with the higher the score, the better. Similarly, the analysis on the collected data will be analyzed

for each research participant to create what is known as the Personal Information Vulnerability Assessment (PIVA) score. This score is comprised of a statistical measurement on all layers and categories, and then calculated to determine a score with a high of 900. The formula is listed below in equation 3-1:

$$PIVA = [P(NC) * W_{NC} + P(PL) * W_{PL} + P(IL) * W_{IL} + P(CL) * W_{CL} + P(S) * W_S] \times 900 \quad (3-1)$$

$P(X)$  represents the probability of X being discovered given a subject's email and phone number and W's represent weighting factors. These abbreviations define the categories of information:

- NC: name and current city
- PL: peripheral layer
- IL: intermediate layer
- CL: core layer
- S: supplemental.

Since each of the category probabilities are conditional probabilities on both email and phone number, they could all be written as,  $P(NC) | P(\text{phone number and email address})$ . In short, the probability of discovering someone's name and current city, depends on the probability of discovering their phone number and email address. However, since the phone number and email address were provided at the beginning for this study, the probabilities are measured as 1, and can be dropped from the final equation and be implied.

Each probability carries a separate weight in the PIVA equation, similar to how each area of credit carries a separate weight with a FICO score. For the PIVA score, the higher the weight, means the associated probability represents a higher vulnerability of personal information. The combination of name and current city has the potential to reveal more information about a research participant than any other category, which is why it carries 35% of the weight. The next highest is the peripheral layer which is weighted at 30%. The third category is the intermediate layer weighted at 15%, followed the core layer at 10% and supplemental at 10%. The weighting factors for PIVA and FICO calculations are shown in Figure 3-6:

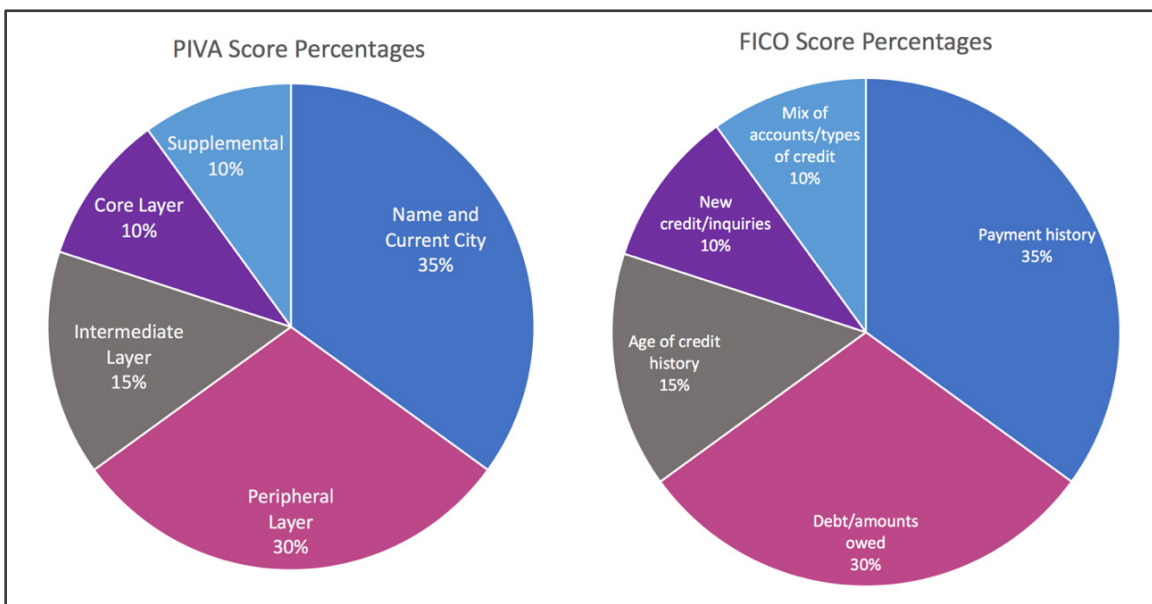


Figure 3-6: FICO and PIVA Weight Distribution

An important difference between the FICO score and the PIVA score is that a high score has different meanings. The FICO score is assessing a person’s credit and a high score indicates excellent credit and is the ideal. The PIVA score is assessing the vulnerability of personal

information. The higher the PIVA score, the more the personal information is vulnerable. For PIVA, a lower score is ideal.

Here is a proof of concept PIVA score calculation. Determining the probabilities for each category comes from the final tally sheet for each research participant. An example tally sheet is shown below in Figure 3-7:

<p><b>Peripheral Layer</b></p> <p>The peripheral layer (also called the identifier layer) is composed of biographic data such as name, age, and gender.</p> <ul style="list-style-type: none"><li><input checked="" type="checkbox"/> Birthday</li><li><input checked="" type="checkbox"/> Name</li><li><input type="checkbox"/> Email Address</li><li><input type="checkbox"/> Phone#</li><li><input checked="" type="checkbox"/> Home Address</li><li><input checked="" type="checkbox"/> Photo</li></ul>
<p><b>Intermediate Layer</b></p> <p>The intermediate layer (also called the demographic layer) is composed of affiliations such as religion, political views, and hometown.</p> <ul style="list-style-type: none"><li><input checked="" type="checkbox"/> Hometown</li><li><input checked="" type="checkbox"/> Current City</li><li><input type="checkbox"/> Religion</li><li><input checked="" type="checkbox"/> Employers</li><li><input checked="" type="checkbox"/> Graduate school/college</li><li><input checked="" type="checkbox"/> High School</li></ul>
<p><b>Core Layer</b></p> <p>The core layer (also known as the favorites and tastes layer) is composed of emotions and values such as favorite music/books, sports, and activities.</p> <ul style="list-style-type: none"><li><input checked="" type="checkbox"/> Favorite Music</li><li><input type="checkbox"/> Favorite Books</li><li><input checked="" type="checkbox"/> Favorite Movies</li><li><input type="checkbox"/> Favorite TV shows</li><li><input type="checkbox"/> Favorite Sports/Teams/Athletes</li><li><input checked="" type="checkbox"/> Hobbies</li></ul>

Figure 3-7: Example Set of Results

Using this example data set, the probability of name and current city  $P(NC) = 1$ , since the name category is ticked in the peripheral layer and current city is ticked in the intermediate layer. The probability of the peripheral layer is equal to the number of discovered categories in the peripheral layer divided by the total number of categories as shown in equation 3-2:

$$P(PL) = \frac{\text{\# of discovered categories}}{\text{total \# of categories}}$$

$$P(PL) = \frac{4}{6} \tag{3-2}$$

$$P(PL) = 0.667$$

The same process is followed for the intermediate and core layers as seen in equation 3-3 and equation 3-4:

$$P(IL) = \frac{\text{\# of discovered categories}}{\text{total \# of categories}}$$

$$P(IL) = \frac{5}{6} \tag{3-3}$$

$$P(IL) = 0.833$$

$$P(CL) = \frac{\text{\# of discovered categories}}{\text{total \# of categories}}$$

$$P(CL) = \frac{3}{6} \tag{3-4}$$

$$P(CL) = 0.50$$

The  $PS(S)$  is the probability that supplemental categories are discovered outside of the defined eighteen. If there are additional categories, then the  $PS(S) = 1$ . For this example, no supplemental categories are included so  $PS(S) = 0$ .

Now that the probabilities are calculated, they can be used in the PIVA equation to get the score:

$$\begin{aligned}
P(NC) &= 1, & P(PL) &= 0.667, & P(IL) &= 0.833, & P(CL) &= 0.50, & P(S) &= 0 \\
PIVA &= [P(NC) * W_{NC} + P(PL) * W_{PL} + P(IL) * W_{IL} + P(CL) * W_{CL} + P(S) * W_S] \times 900 \\
PIVA &= [(1)(0.35) + (0.667)(0.30) + (0.833)(0.15) + (0.50)(0.10) + (0)(0.10)] \times 900 \\
PIVA &= [0.35 + 0.20 + 0.125 + 0.05 + 0] \times 900 \\
PIVA &= 0.725 \times 900 \\
PIVA \text{ Score} &= 653
\end{aligned}
\tag{3-5}$$

The PIVA Score is 653. When all of the PIVA Scores are calculated for each research participant, then they can be arranged in a normal distribution by calculating the average and standard deviation. The standard deviation calculation will use the following formula in equation 3-4:

$$\sqrt{\frac{\sum(x - \bar{x})^2}{(n - 1)}}
\tag{3-6}$$

Normal distribution will determine where each score falls in relation to all scores, and make it possible to determine which scores are average, high and low.



## **4 DIGITAL FOOTPRINT COLLECTION AND ANALYSIS**

The digital footprint collection consisted of gathering information from seven different participants. This chapter goes into detail on the research of these participants.

One of the challenges with researching digital footprints is that no two footprints are exactly the same. Scoping this project was a balance between acquiring enough data from each research subject for a good analysis, while not exhausting available resources. The predetermined list of categories was meant to be flexible in case the research became too cumbersome. Thankfully, it turned out to be an ideal balance.

### **4.1 Research Results**

Research participants are assigned IDs of RP\_A through RP\_H with ‘RP’ standing for research participant. The final results for all research participants are located in Appendix A. The blurred lines indicate redacted information to protect the participant. Below is a sample result for RP\_G in Figure 4-1 and Figure 4-2:

The data for this research participant was acquired through a wide variety of sources as seen by the difference in URLs. What’s interesting is this research participant’s phone number showed as “unlisted” for most of the search results. Normally, the phone number was one of the most revealing identifiers for other research participants, but in this case, the results came primarily from the supplemental phase of the research. The following sections go into detail on

the analysis for the identifiers, categories, and participant results. Measuring effectiveness for each of the four research phases is suggested in chapter 5 as part of future work.

**ID: RP\_G**

**Peripheral Layer**  
 The peripheral layer (also called the identifier layer) is composed of biographic data such as name, age, and gender.

- Birthday
- Name  
 Possible full name? Looks like Laurel might be a middle name  
[https://registrar. \[redacted\] /registrar/graduation/commencement \[redacted\]](https://registrar. [redacted] /registrar/graduation/commencement [redacted])  
[https://pipl.com/search?q= \[redacted\] t&l=&sloc=&in=6](https://pipl.com/search?q= [redacted] t&l=&sloc=&in=6)
- Found middle name
- Email Address  
[https://pipl.com/search?q= \[redacted\] t&l=&sloc=&in=6](https://pipl.com/search?q= [redacted] t&l=&sloc=&in=6)
- Other Phone#  
[https://www.truepeoplesearch.com/results?name= \[redacted\] &rid=0x0](https://www.truepeoplesearch.com/results?name= [redacted] &rid=0x0)
- Home Address  
[http://www.homefacts.com/address/Illinois/Mchenry-County/ \[redacted\] : \[redacted\]](http://www.homefacts.com/address/Illinois/Mchenry-County/ [redacted] : [redacted])
- Photo  
[https://pipl.com/search?q= \[redacted\] &l=&sloc=&in=6](https://pipl.com/search?q= [redacted] &l=&sloc=&in=6)  
[https://www.facebook.com/ \[redacted\]](https://www.facebook.com/ [redacted])

**Intermediate Layer**  
 The intermediate layer (also called the demographic layer) is composed of affiliations such as religion, political views, and hometown.

- Hometown  
[https://www.facebook.com/ \[redacted\]](https://www.facebook.com/ [redacted])  
[http://www.switchboard.com/phone/1- \[redacted\]](http://www.switchboard.com/phone/1- [redacted])
- Current City  
[http://www.homefacts.com/address/Illinois/Mchenry-County/ \[redacted\] : \[redacted\]](http://www.homefacts.com/address/Illinois/Mchenry-County/ [redacted] : [redacted])  
[Intelius](http://www.intelius.com/)  
 [Maltego] [https://plus.google.com/116799 \[redacted\]](https://plus.google.com/116799 [redacted])
- Religion  
[https://www.facebook.com/ \[redacted\] about?section=contact-info&pnref=about](https://www.facebook.com/ [redacted] about?section=contact-info&pnref=about)  
[https://www.pandora.com/profile/thumbs/ \[redacted\]](https://www.pandora.com/profile/thumbs/ [redacted])
- Employers
- Graduate school/college  
[https://registrar. \[redacted\] /registrar/graduation/commencement \[redacted\]](https://registrar. [redacted] /registrar/graduation/commencement [redacted])  
[https://scontent-sjc2-1.xx.fbcdn.net/v/ \[redacted\] oh=d89 \[redacted\] 1b1&oe=58963853](https://scontent-sjc2-1.xx.fbcdn.net/v/ [redacted] oh=d89 [redacted] 1b1&oe=58963853)  
 [Maltego] [https://plus.google.com, \[redacted\]](https://plus.google.com, [redacted])
- High School

Figure 4-1: RP\_G Results

### Core Layer

The core layer (also known as the favorites and tastes layer) is composed of emotions and values such as favorite music/books, sports, and activities.

- Favorite Music  
[Maltego] <https://www.pandora.com/profile/>  
<https://www.pandora.com/profile/thumbs/>  
<https://www.spokeo.com/social/profile?q=sw-xZ2AydC...YFomlKA2cc&loaded=1>
- Favorite Books  
[https://www.facebook.com/.../photos?source\\_ref=pb\\_friends\\_tl](https://www.facebook.com/.../photos?source_ref=pb_friends_tl)
- Favorite Movies  
[https://www.facebook.com/.../photos?source\\_ref=pb\\_friends\\_tl](https://www.facebook.com/.../photos?source_ref=pb_friends_tl)
- Favorite TV shows  
[https://www.facebook.com/.../photos?source\\_ref=pb\\_friends\\_tl](https://www.facebook.com/.../photos?source_ref=pb_friends_tl)
- Favorite Sports/Teams/Athletes
- Hobbies  
[https://registrar.byu.edu/registrar/graduation/commencement/2015Su\\_Program.pdf](https://registrar.byu.edu/registrar/graduation/commencement/2015Su_Program.pdf)  
<https://www.pinterest.com/>  
Possible pinterest page  
<https://www.facebook.com/>

### Other Info

- Relatives/Associates  
<http://www.homefacts.com/address/Illinois/Mchenry-County/Crystal-Lake/60014/>  
<https://www.truepeoplesearch.com/results?name=&rid=0x0>  
[Maltego] <https://plus.google.com/116...475>
- Home Value  
<http://www.homefacts.com/address/Illinois/Mchenry-County/Crystal-Lake/60014/>
- Sale information on their house  
<https://www.co.mchenry.il.us/home/showdocument?id=...>
- Wedding month/year  
[https://scontent-sjc2-1.xx.fbcdn.net/v/t1.0-9/549164\\_550420261654997\\_1441602441\\_n.jpg?oh=97160ada45c666e1...=588D197F](https://scontent-sjc2-1.xx.fbcdn.net/v/t1.0-9/549164_550420261654997_1441602441_n.jpg?oh=97160ada45c666e1...=588D197F)
- Spouse  
[https://scontent-sjc2-1.xx.fbcdn.net/t31.0-8/10355535\\_78...\\_1306614173807182393\\_o.jpg](https://scontent-sjc2-1.xx.fbcdn.net/t31.0-8/10355535_78..._1306614173807182393_o.jpg)  
Intelius
- Political views  
<https://www.facebook.com/>

Figure 4-2: RP\_G Results Continued

## 4.2 Analysis of Identifiers

The research pivoted on two main identifiers, email address and phone number. The project hypothesized that these two identifiers would be successful in uncovering further information,

however, the efficacy of each identifier was unknown in advance. The success rates given an email and phone number are listed in Table 4-1:

Table 4-1: Success Rates for Layers Given Email and Phone Number

Layer	Success Rate
Peripheral	77.08%
Intermediate	64.58%
Core	54.17%
Overall	65.28%

The results indicate that of the 18 categories, with a phone number and email address an average of 11.75 could be discovered. That gives an overall success rate of 65.28%. Note that the peripheral layer, which contains the information that is most unique to an individual, had the highest success rate of 77.08%. Therefore, if provided with an email address and phone number, a researcher will be 1.2 to 1.4 times more successful in finding uniquely identifying information than demographic or preferential information. This disproves the assumption that highly personal information is more difficult to find.

These success calculations are based on having both an email address and phone number since that is how the data was collected and organized for all of the research participants. In order to calculate a separate success rate for an email address and phone number, the research was revisited for RP\_A and RP\_H. All of the data for RP\_A was revisited to determine what information was discovered with only a phone number, and RP\_H's data was revisited to determine what information was discovered with only an email address. The results are below in Figure 4-3 and Figure 4-4:

## RP\_A: Phone Number Data

### Peripheral Layer

The peripheral layer (also called the identifier layer) is composed of biographic data such as name, age, and gender.

- Birthday
- Name
- Email Address
- Cell Phone#
- Other Phone#
- Home Address
- Photo

### Intermediate Layer

The intermediate layer (also called the demographic layer) is composed of affiliations such as religion, political views, and hometown.

- Hometown
- Current City
- Religion
- Employers
- Graduate school/college
- High School

### Core Layer

The core layer (also known as the favorites and tastes layer) is composed of emotions and values such as favorite music/books, sports, and activities.

- Favorite Music
- Favorite Books
- Favorite Movies
- Favorite TV shows
- Favorite Sports/Teams/Athletes
- Hobbies

### Other Info

- Relatives
- Mother's Maiden name
- Phone Type

Figure 4-3: Research Participant RP\_A Results Using Only Phone Number

## RP\_H: Email Address Data

**Peripheral Layer**  
 The peripheral layer (also called the identifier layer) is composed of biographic data such as name, age, and gender.

- Birthday
- Name
- Email Address
- Cell Phone#
- Other Phone#
- Home Address
- Photo

**Intermediate Layer**  
 The intermediate layer (also called the demographic layer) is composed of affiliations such as religion, political views, and hometown.

- Hometown
- Current City
- Religion
- Employers
- Graduate school/college
- High School

**Core Layer**  
 The core layer (also known as the favorites and tastes layer) is composed of emotions and values such as favorite music/books, sports, and activities.

- Favorite Music
- Favorite Books
- Favorite Movies
- Favorite TV shows
- Favorite Sports/Teams/Athletes
- Hobbies

Figure 4-4: Research Participant RP\_H Results Using Only Email Address

RP\_A and RP\_H were selected because each of them had the highest amount of information uncovered compared to the rest of the research participants and this proof of concept will encapsulate a worst-case scenario for both the phone number and email address respectively. Figure 4-5 is proof of concept for what is possible to gain with each identifier separately.

In this instance, an email address is capable of disclosing information from each of the three layers: peripheral, intermediate and core. It's most effective with the intermediate layer, but overall captures nearly half of the 18 categories.

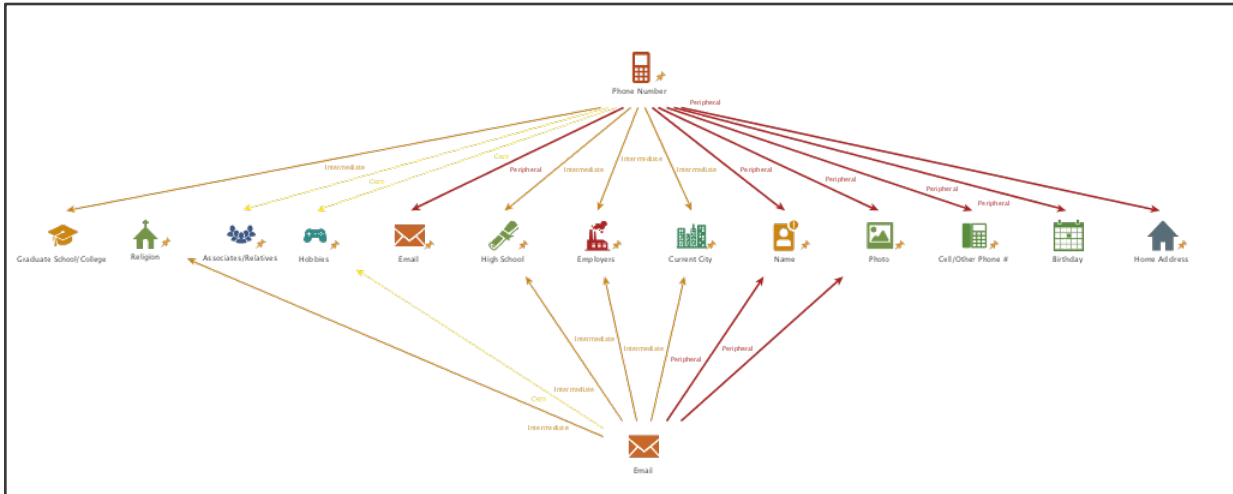


Figure 4-5: Categories Discovered from Phone and Email Separately

The phone number revealed 12 of the 18 categories, including all categories in the peripheral layer. This demonstrates that each identifier can reveal multiple levels of information. The phone number in this instance has a greater impact by 4 categories.

Figure 4-6 shows how categories are mapped to an email address or phone number, or in some cases, both. Some map to only one, like favorite books, while others map to both, like current city. There is a total of seven categories that map to both an email address and phone number.

Figure 4-7 displays the same information, but each node is weighted based on how many connections that node has. The bigger nodes have multiple connections showing that they are capable of revealing a variety of information. This analysis of the phone number and email address primarily highlight the fact that they are both influential identifiers in uncovering personal information from all layers and are sufficient to uncover the needed data for each research participant in this study.

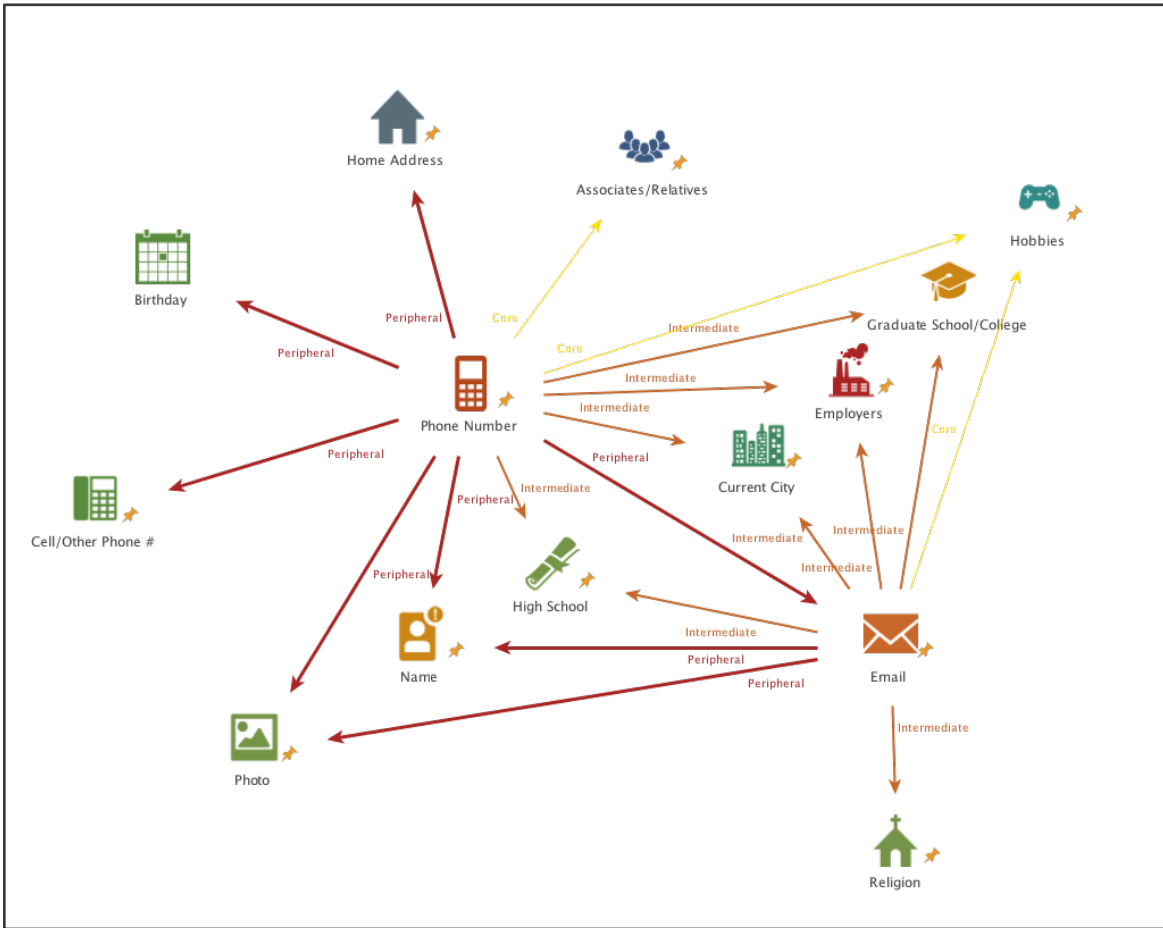


Figure 4-6: Mapping of Categories to Email Address and Phone Number



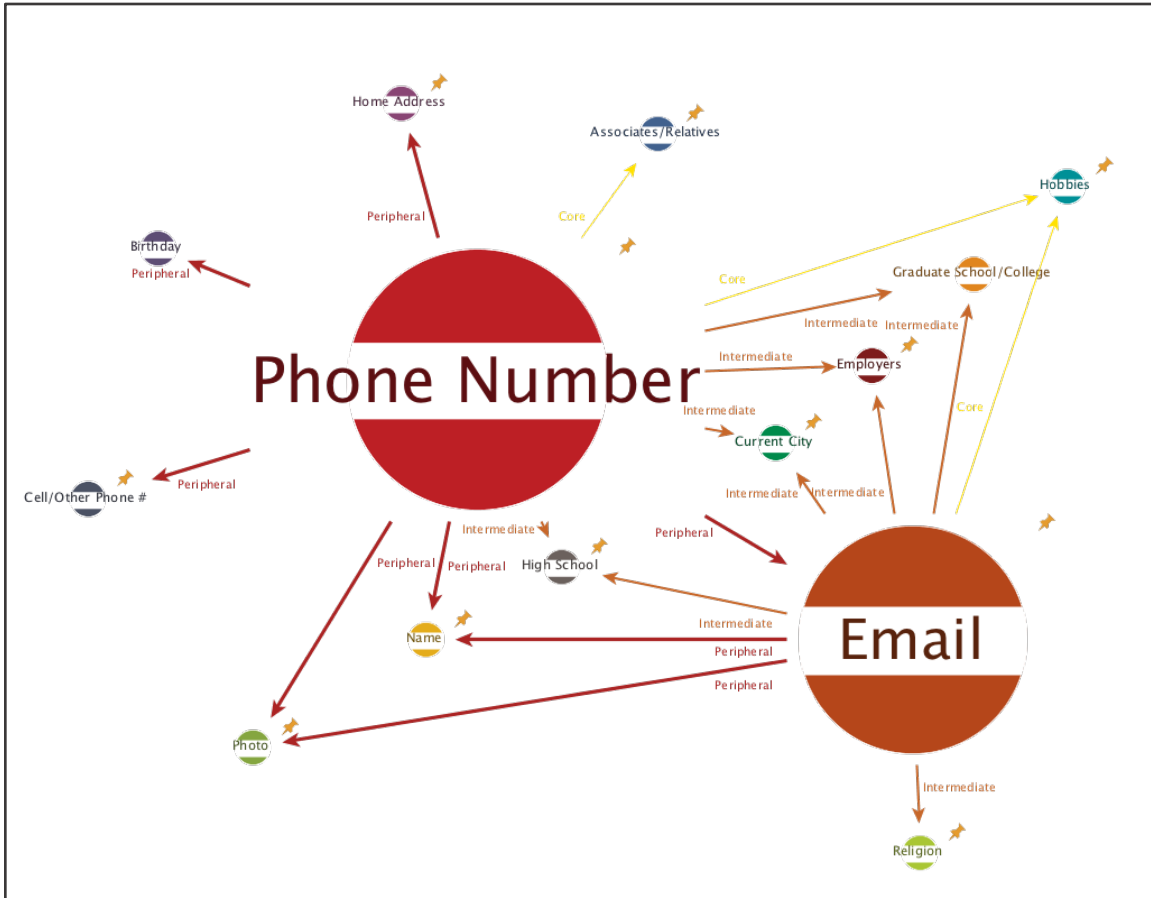


Figure 4-7: Weighted Mapping of Categories to Phone and Email

### 4.3 Analysis of Categories

Calculating the probability for discovering a specified category is shown in equation 4-1:

$$P(\text{Category}) | P(\text{Email and Phone number}) = \frac{\sum_{i=RP\_A}^{RP\_H} \text{Category discovered}}{\text{Total number of research participants}} \quad (4-1)$$

Again, like the PIVA Score calculation, this is a conditional probability dependent on the probability of obtaining an email address and phone number. Since the probability for obtaining an email address and phone number in this scenario is 1, it will be dropped from the equation and

implied. As an example, the category birthdate, was discovered 5 times among 8 research participants. Plugging those numbers into equation 4-1 would be:

$$P(\textit{Birthdate}) = \frac{5}{8} \quad (4-2)$$

$$P(\textit{Birthdate}) = 0.625$$

In short, the probability of discovering a birthdate is 62.5%. The probabilities for each of the categories are shown in Figure 4-8. All categories in the peripheral layer had probabilities greater than 60%, whereas the intermediate and core layers had a wider range from 37% to 100%. This indicates that an email address and phone number are more successful in uncovering information in the peripheral layer than any other layer and likely represents the amount of information publicly posted online.

Current city and hobbies both have probabilities of 1, meaning those categories were discovered for every single research participant. Recall that the heaviest weighted category of the PIVA score is 35% for name and current city. Since the probability for discovering the current city is 100% and the name is 87.5%, all participants will have it as part of their PIVA score.

One measurement that could not be captured very well is the quantity of data for each category. Categories in the peripheral layer typically have one piece of data per category, whereas categories in the intermediate and especially core can have multiple pieces of data per category. Although the quantity couldn't be measured, quality was not affected. For a category such as hobbies, only one is necessary to be effective in something like a phishing campaign.

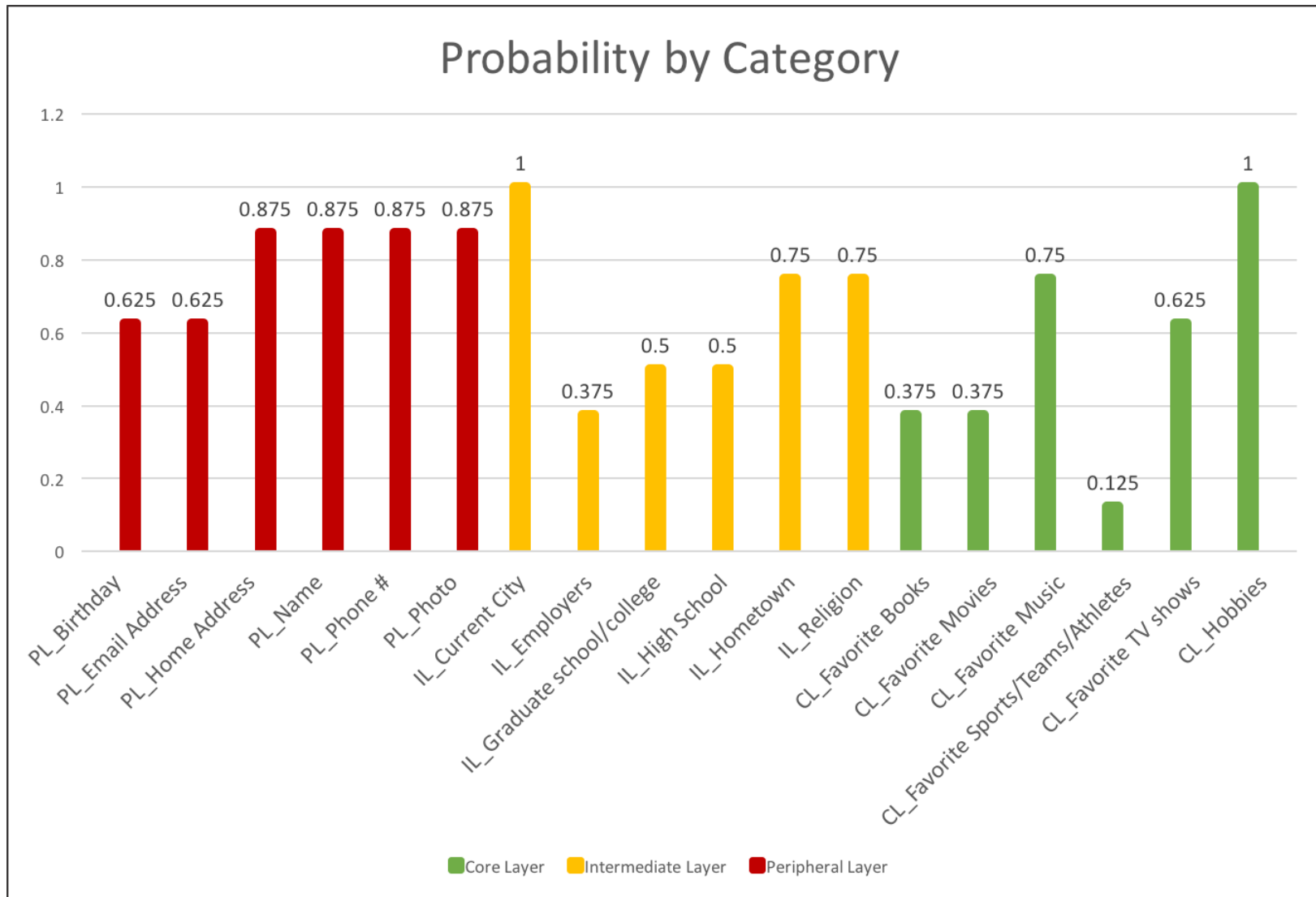


Figure 4-8: Probability by Category

Another element that could affect the probability results is the difficulty of redacting information. Most categories found in the core layer are often found on sites such as social media, whereas information in the intermediate and peripheral layers are found within public records. It's possible core layer information could be harder to find because it's easier to redact.

#### 4.4 Analysis of the Research Participants: PIVA Score

The PIVA score calculation was developed for this project to assess the vulnerability of personal information for each research participant in a way that captures information discoverability at varying levels of information value.

Table 4-2 shows the PIVA score for each research participant:

Table 4-2: PIVA Scores for Research Participants

Participant ID	PIVA Score
RP_A	833
RP_B	720
RP_D	758
RP_E	750
RP_F	750
RP_G	795
RP_H	833

The majority of scores are in the 700-800 range. The normal distribution calculations of the remaining PIVA scores are shown in Figure 4-9. The average is 777 with a standard deviation of 44.  $\pm 1$  standard deviation is 733-821 which has an *area(probability)* = 0.683. This will become the standard to which the PIVA score is rated on the graph.

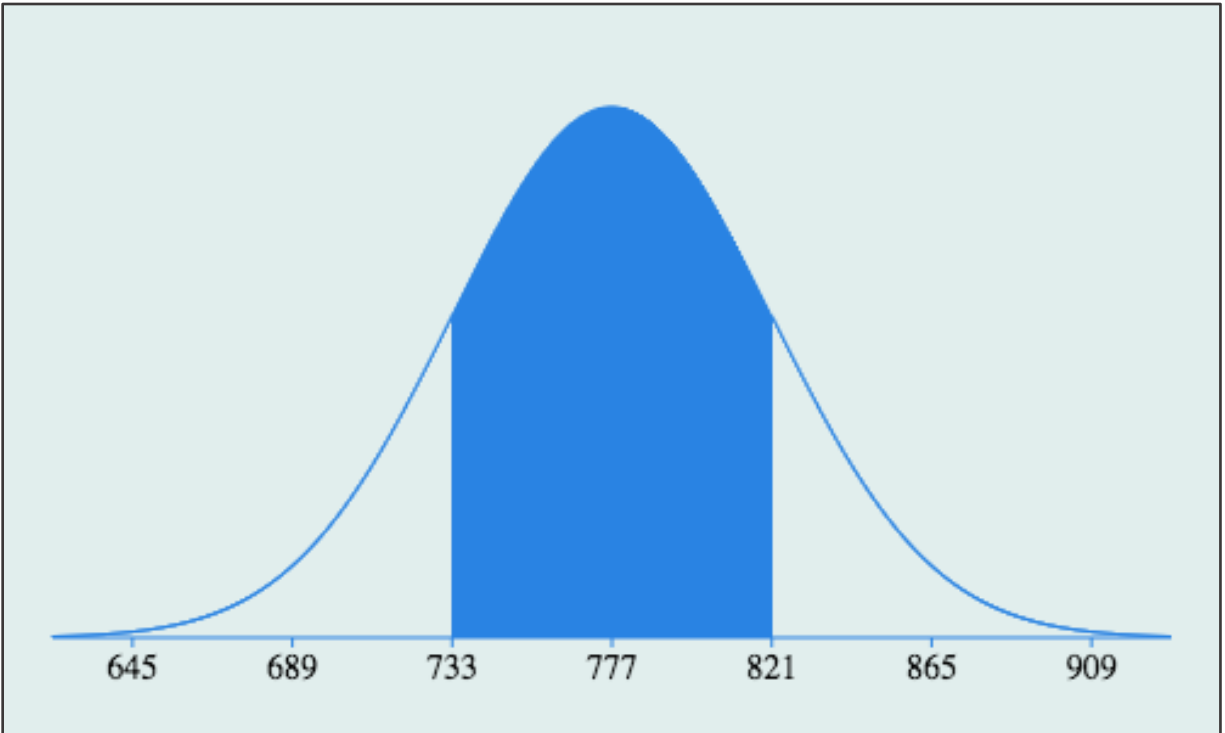


Figure 4-9: Normal Distribution of PIVA Scores

If the calculated PIVA score falls within the range of 733-821 it will be considered average and marked yellow. If the calculated PIVA score is higher than 821, then it will be marked red, indicating that the information vulnerability is higher than normal. The ultimate goal is to have a lower score, so any scores calculated at less than 733 will be marked green. Figure 4-10 shows the PIVA score for RP\_A.

This PIVA score is the record high at 833. It is marked red since it's greater than the average range of 733-821 and indicates large amounts of exposed information across all sensitivity layers.

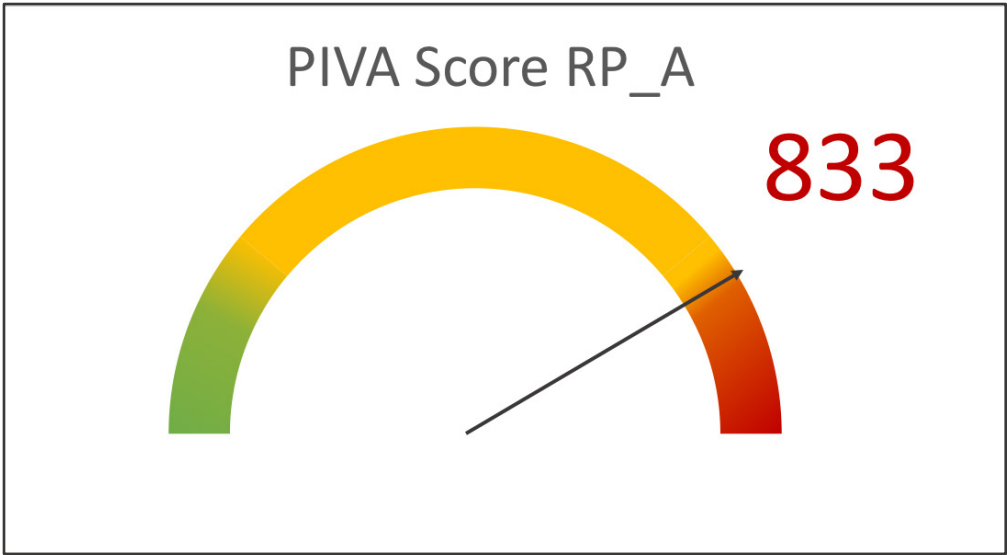


Figure 4-10: PIVA Score for RP\_A

## 5 CONCLUSIONS AND FUTURE WORK

The purpose of this research was to educate and inform users on the breadth and depth of their digital footprint by addressing two different questions and hypotheses:

- (R1) How much information can be uncovered about a person given one or two details?
- (R2) How can a digital footprint be captured and measured in size and impact using various metrics?
- Hypothesis (H1): Details that are generally unique to an individual yet common (e.g. phone number and email address) reveal personal information. Because these details are used in association with areas like work, school, home, and hobbies, it's expected they would lead to information that will reveal a great deal about each of those areas.

From the research, it was found that an email address and phone number were sufficient to uncover the breadth of information exposure for an individual. Also, the PIVA calculation was formulated to assess the impact of information exposure and report the greatest areas of risk based on a weighted scale.

## 5.1 Project Contributions

This project provides the following contributions:

- Proved email addresses and phone numbers are effective in discovering personal information regardless of demographic.
- Proved information characterized as highly personal is 1.2 to 1.4 times more likely to be discovered.
- Provided a metric known as the PIVA score that captures the vulnerability and extent of exposed personal information based on research.
- Documented the process and results of ethical and secure Open-Source Intelligence (OSINT) gathering on seven research participants.

Information disclosure affects all users online. Data collection is invisible, privacy notices are difficult to understand, and sensitive data regarding health, finances, or children are collected for behavioral advertising. The majority of users show concern for information privacy, but few have adopted protective behaviors. The key is knowing where to start.

Research concluded that personal information is readily available online and easily acquired, regardless of age or demographic. Protecting personal information should begin with identifiers such as email addresses and phone numbers, since they were proven to reveal information across all levels of sensitivity.

This project allows users to understand exactly what types of information are exposed, and how they can be discovered. No other research has attempted to extensively research individuals using a small number of identifiers and analyze the resulting information exposure. Furthermore, this project has adopted and modified the FICO metric to create a new PIVA metric



that encapsulates the vulnerability and extent of exposed information to educate users and provide insight into areas of greatest risk.

## 5.2 Future Research

This research lends itself to a variety of future research projects:

- Generate separate result lists for each phase of research (instead of a combined final result) to analyze the effectiveness of each phase.
- Study each resource individually to determine which tools are most effective.
- Research the probability of acquiring identifiers such as email addresses and phone numbers since the current study used a probability of 1.
- Repeat the research using different identifiers.
- Repeat the research using additional or alternate categories.
- Repeat the research using additional participants with more diverse demographics.
- Research the effectiveness of personal information redaction techniques by performing a before and after analysis of a research participant that used the techniques.
- Include before and after surveys that determine if PIVA scores are an effective tool to raise user awareness about personal information vulnerability.
- Expand the list of resources, or develop a tool to acquire personal information.
- Improve PIVA score formula.

Information privacy is a growing concern and there are plenty of areas that are yet to be explored, as seen from this list. A logical next step would be a study of the adjustment of a PIVA score before and after remediation efforts. The design behind the PIVA score was to provide a point in time metric of information exposure that individuals could aim to improve. If an individual undergoes remediation efforts to better protect their information and their PIVA score improves, then it would demonstrate that the metric works as designed.

Another current limitation of the PIVA score is that the categories of name and current city are counted twice: once as part of their own category, and then a second time as part of the peripheral and intermediate layer probabilities. The use of additional categories would make it possible for the ratios of category to layer to remain the same without double counting the name and current city categories in the PIVA equation.

Additionally, the PIVA score could consist of various equations that each represent various scenarios. For instance, one scenario could be the vulnerability of social engineering attacks (e.g. spear phishing). This type of vulnerability would put greater emphasis on the intermediate and core layers since that type of information is generally used for social engineering. Another view could be the vulnerability for identity theft. This type of vulnerability would put greater emphasis on the peripheral layer or applicable supplemental data, since oftentimes password reset or security questions often include information found within those areas.

### **5.3 Potential Applications**

This project was designed to apply to users outside of the study. The PIVA scores from the participants in the study can be used by others to start protecting their own information by

understanding common pitfalls from participants in the study. If the research could be automated, it could be applied to a larger population as a service. Just as financial institutions utilize FICO scores to determine interest rates, they could also utilize PIVA scores. A low PIVA score indicates a low probability of information vulnerability which also means a lower probability of identity fraud. Financial institutions could incentivize customers who show dedication toward protecting personal information and maintaining a low PIVA score.

Information privacy is a growing concern and more are taking advantage of it rather than protecting it. This project is one step toward changing that.

## REFERENCES

- Activities, BYU Office of Research and Creative. 2017. "Irb | Office of Research & Creative Activities." <https://orca.byu.edu/irb/>.
- Bartel, S. 1999. "An Investigation of Gender Differences in on-Line Privacy Concerns and Resultant Behaviors." *Journal of Interactive Marketing* 13, no. 4.
- Dienlin, T., and Metzger, M. 2016. "An Extended Privacy Calculus Model for Snss: Analyzing Self-Disclosure and Self-Withdrawal in a Representative U.S. Sample." *Journal of Computer-Mediated Communication* 21, no. 5 (2016).
- Dienlin, T., and Trepte, S. 2015. "Is the Privacy Paradox a Relic of the Past? An in-Depth Analysis of Privacy Attitudes and Privacy Behaviors." (2015).
- Irani, D., Webb, S., Li, K., and Pu, C. 2009. "Large Online Social Footprints - an Emerging Threat." *In Computational Science and Engineering, 2009. CSE'09. International Conference 3*: 271-276.
- Kokolakis, S. 2017. "Privacy Attitudes and Privacy Behaviour: A Review of Current Research on the Privacy Paradox Phenomenon." *Computers & Security* 64, no. Supplement C (2017/01/01/): 122-134.
- Krishnamurthy, B., Naryshkin, K., and Wills, C. 2011. "Privacy Leakage Vs. Protection Measures: The Growing Disconnect." *Proceedings of the Web 2*: 1-10.
- Lee, S., Lee, Y., Lee, J., and Park, J. 2015. "Personalized E-Services: Consumer Privacy Concern and Information Sharing." *Social Behavior and Personality: an international journal* 43, no. 5: 729-740.
- Malandrino, D., Petta, A., Scarano, V., Serra, L., Spinelli, R., and Krishnamurthy, B. 2013. "Privacy Awareness About Information Leakage: Who Knows What About Me?" *Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society*: 279-284.
- Malhotra, A., Totti, L., Meira, W., Kumaraguru, P., and Almeida, V. 2012. "Studying User Footprints in Different Online Social Networks." 1065-1070.
- Matzner, T., Masur, P., Ochs, C., and Pape, T. 2017. "Do-It-Yourself Data Protection - Empowerment or Burden?" *In Data Protection on the Move*, 277-305: Springer Netherlands.

- Maxwell, W., and Teng, C. 2016. *Cinderella or Big Foot? Analysing the Size and Impact of Digital Footprints. European Conference on Social Media*. Caen, France: Academic Conferences and Publishing International Limited.
- Norberg, P., Horne, D., and Horne, D. 2007. "The Privacy Paradox: Personal Information Disclosure Intentions Versus Behaviors." *Journal of Consumer Affairs*: 100-126.
- Otsuki, M., and Sonehara, N. 2013. "Estimating the Value of Personal Information with Sns Utility." 512-516.
- Rose, E. 2005. "Data Users Versus Data Subjects Are Consumers Willing to Pay for Property Rights to Personal Information." *InSystem Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference*.
- Schneier, B. 2015. *Data and Goliath : The Hidden Battles to Collect Your Data and Control Your World*. First edition. ed.
- Shi, X., Li, D., Zhu, H., and Zhang, W. 2007. "Research on Supply Chain Information Classification Based on Information Value and Information Sensitivity." *InService Systems and Service Management, 2007 International Conference*: 1-7.
- Shibchurn, J., and Xiang Bin, V. 2014. "Investigating Effects of Monetary Reward on Information Disclosure by Online Social Networks Users." 1725-1734.
- Ur, B., Leon, P., Cranor, L., Shay, R., and Wang, Y. 2012. *Smart, Useful, Scary, Creepy: Perceptions of Online Behavioral Advertising. Proceedings of the Eighth Symposium on Usable Privacy and Security*: ACM.
- Wisniewski, P., Knijnenburg, B., and Lipford, H. 2014. "Profiling Facebook Users' Privacy Behaviors." *SOUPS2014 Workshop on Privacy Personas and Segmentation*.

## APPENDIX A. RESEARCH RESULTS

ID: RP\_A

### Peripheral Layer

The peripheral layer (also called the identifier layer) is composed of biographic data such as name, age, and gender.

- Birthday  
[Intelius](#)
- Name  
Gmail
- Email Address  
[Instantcheckmate](#)
- Cell Phone#  
[Intelius](#)  
[Instantcheckmate](#)
- Other Phone#  
[https://pipl.com/search?q=\[REDACTED\]&l=&sloc=&in=6](https://pipl.com/search?q=[REDACTED]&l=&sloc=&in=6)  
[https://www.truepeoplesearch.com/results?name=\[REDACTED\]&rid=0xll](https://www.truepeoplesearch.com/results?name=[REDACTED]&rid=0xll)  
[Instantcheckmate](#)
- Home Address  
[Intelius](#)  
[https://www.truepeoplesearch.com/results?name=\[REDACTED\]&rid=0xll](https://www.truepeoplesearch.com/results?name=[REDACTED]&rid=0xll)  
[Instantcheckmate](#)
- Photo  
[Intelius](#)  
[Instantcheckmate](#)

### Intermediate Layer

The intermediate layer (also called the demographic layer) is composed of affiliations such as religion, political views, and hometown.

- Hometown  
[https://www.facebook.com/\[REDACTED\]](https://www.facebook.com/[REDACTED])  
[Instantcheckmate](#)
- Current City  
[https://www.truepeoplesearch.com/results?name=\[REDACTED\]&rid=0xll](https://www.truepeoplesearch.com/results?name=[REDACTED]&rid=0xll)  
[Instantcheckmate](#)  
[Instantcheckmate](#)
- Religion  
[https://www.truepeoplesearch.com/results?name=\[REDACTED\]&rid=0xll](https://www.truepeoplesearch.com/results?name=[REDACTED]&rid=0xll)
- Employers  
[https://www.truepeoplesearch.com/results?name=\[REDACTED\]&rid=0xll](https://www.truepeoplesearch.com/results?name=[REDACTED]&rid=0xll)  
[Instantcheckmate](#)
- Graduate school/college  
[Instantcheckmate](#)
- High School

### Core Layer

The core layer (also known as the favorites and tastes layer) is composed of emotions and values such as favorite music/books, sports, and activities.

- Favorite Music  
<https://www.facebook.com/...>
- Favorite Books
- Favorite Movies
- Favorite TV shows
- Favorite Sports/Teams/Athletes  
<https://www.facebook.com/...>
- Hobbies  
<https://www.truepeoplesearch.com/results?name=...&rid=0xll>

### Additional Info

- Spouse  
[Intelius](https://www.truepeoplesearch.com/results?name=...&rid=0xll)
- Acquaintances  
[Intelius](https://www.truepeoplesearch.com/results?name=...&rid=0xll)
- Car VIN #  
<http://www.switchboard.com/phone/...>
- Verified that Email was ~~pwned~~ in ~~HavelBeenPwned~~  
<http://haveibeenpwned>
- Myspace Account  
<http://haveibeenpwned>
- Middle Name  
[Intelius](https://www.truepeoplesearch.com/results?name=...&rid=0xll)
- College Degree  
[Intelius](https://www.truepeoplesearch.com/results?name=...&rid=0xll)
- Age  
<https://www.truepeoplesearch.com/results?name=...&rid=0xll>
- LinkedIn  
[Instantcheckmate](https://www.truepeoplesearch.com/results?name=...&rid=0xll)
- Facebook  
[Instantcheckmate](https://www.truepeoplesearch.com/results?name=...&rid=0xll)

ID: RP\_B

### Peripheral Layer

The peripheral layer (also called the identifier layer) is composed of biographic data such as name, age, and gender.

- Birthday
- Name  
Gmail
- Email Address  
<http://www.switchboard.com/phone/1->
- Other Phone#
- Home Address  
<http://www.switchboard.com/phone/1->  
<https://nuumber.com/person/563a2b35>
- Photo  
<https://www.pinterest.com/>

### Intermediate Layer

The intermediate layer (also called the demographic layer) is composed of affiliations such as religion, political views, and hometown.

- Hometown  
<http://ww3.org/clc/studentervices/Pages/Home.aspx>
- Current City  
<https://nuumber.com/person/563a2b35cf>
- Religion  
<https://www.pinterest.com/>
- Employers
- Graduate school/college
- High School  
<http://ww3.org/clc/studentervices/Pages/Home.aspx>

### Core Layer

The core layer (also known as the favorites and tastes layer) is composed of emotions and values such as favorite music/books, sports, and activities.

- Favorite Music (try and find Pandora account?)  
<http://www.pandora.com/people/>
- Favorite Books
- Favorite Movies
- Favorite TV shows  
<https://www.pinterest.com/>
- Favorite Sports/Teams/Athletes
- Hobbies  
<https://www.pinterest.com/>

### Other Info

- Relatives/Associates  
<https://nuumber.com/person/563a2b35>  
<https://nuumber.com/person/563a2b35>  
<http://hodges-directory.us/directory.php?q=3o52-8>  
[Intelius](http://www.intelius.com/)
- Mother's Maiden Name  
[Whitenpages](http://www.whitenpages.com/)



ID: RP\_C

### Peripheral Layer

The peripheral layer (also called the identifier layer) is composed of biographic data such as name, age, and gender.

- Birthday
- Name
- Email Address
- Other Phone#  
<http://www.refugeforums.com/refuge/threads/british-lab-...>  
[Whitpages Number](#)
- Home Address  
[Whitpages Number](#)
- Photo

### Intermediate Layer

The intermediate layer (also called the demographic layer) is composed of affiliations such as religion, political views, and hometown.

- Hometown  
<http://nprn.org/404/376/2329/40...html>  
<http://www.refugeforums.com/refuge/threads/british-lab-...>
- Current City  
[Whitpages Number](#)
- Religion
- Employers
- Graduate school/college
- High School

### Core Layer

The core layer (also known as the favorites and tastes layer) is composed of emotions and values such as favorite music/books, sports, and activities.

- Favorite Music
- Favorite Books
- Favorite Movies
- Favorite TV shows
- Favorite Sports/Teams/Athletes
- Hobbies  
<http://www.refugeforums.com/refuge/threads/british-lab-...>

### Other Info

- Relatives  
<http://www.refugeforums.com/refuge/threads/british-lab-.../>  
[Whitpages Number](#)
- Mother's Maiden name  
<https://www.linkedin.com/in/...>  
[Whitpages Number](#)
- Possible Namesake  
<https://www.truepeoplesearch.com/results?name=...> 0x0

ID: RP\_D

### Peripheral Layer

The peripheral layer (also called the identifier layer) is composed of biographic data such as name, age, and gender.

- Birthday  
[Intelius](#)
- Name  
Gmail
- Email Address  
[Intelius](#)
- Other Phone#  
[Intelius](#)
- Home Address  
[Intelius](#)
- Photo

### Intermediate Layer

The intermediate layer (also called the demographic layer) is composed of affiliations such as religion, political views, and hometown.

- Hometown
- Current City  
[Whitenpages](#)
- Religion
- Employers  
[Intelius](#)
- Graduate school/college  
[Intelius](#)
- High School

### Core Layer

The core layer (also known as the favorites and tastes layer) is composed of emotions and values such as favorite music/books, sports, and activities.

- Favorite Music
- Favorite Books  
[Intelius](#)
- Favorite Movies  
[Intelius](#)
- Favorite TV shows  
[Intelius](#)
- Favorite Sports/Teams/Athletes
- Hobbies  
[Intelius](#)

### Extra Info Found

- Relatives and Relative Work History  
[Intelius](#)
- Pet/Pet Name  
[Intelius](#)
- Phone Carrier  
[Whitenpages](#)
- Maiden Name  
[Intelius](#)

ID: RP\_E

Peripheral Layer

The peripheral layer (also called the identifier layer) is composed of biographic data such as name, age, and gender.

- Birthday
- Name
- Gmail
- Email Address
- Other Phone#
- [Intelius](#)
- Home Address  
<https://nuwber.com/person/563a2b>
- Photo  
<https://www.pinterest.com>

Intermediate Layer

The intermediate layer (also called the demographic layer) is composed of affiliations such as religion, political views, and hometown.

- Hometown  
<https://nuwber.com/person/563a2b39cf>
- Current City  
<https://nuwber.com/person/563a2b39cf>
- Religion  
<https://www.facebook.com/>  
<https://www.pinterest.com/> /byutv/
- Employers
- Graduate school/college
- High School  
<http://home.d47.org/bastrebel/files/> pdf

### Core Layer

The core layer (also known as the favorites and tastes layer) is composed of emotions and values such as favorite music/books, sports, and activities.

- Favorite Music  
[https://www.pinterest.com/\[REDACTED\]/hamilton/](https://www.pinterest.com/[REDACTED]/hamilton/)
- Favorite Books  
[https://www.pinterest.com/\[REDACTED\]/books/](https://www.pinterest.com/[REDACTED]/books/)  
[https://www.pinterest.com/\[REDACTED\]/harry-potter/](https://www.pinterest.com/[REDACTED]/harry-potter/)
- Favorite Movies  
[https://www.pinterest.com/\[REDACTED\]/star-wars/](https://www.pinterest.com/[REDACTED]/star-wars/)  
[https://www.pinterest.com/\[REDACTED\]/marvel/](https://www.pinterest.com/[REDACTED]/marvel/)  
[https://www.pinterest.com/\[REDACTED\]/disney/](https://www.pinterest.com/[REDACTED]/disney/)  
[https://www.pinterest.com/\[REDACTED\]/harry-potter/](https://www.pinterest.com/[REDACTED]/harry-potter/)
- Favorite TV shows  
[https://www.pinterest.com/\[REDACTED\]/byutv/](https://www.pinterest.com/[REDACTED]/byutv/)  
[https://www.pinterest.com/\[REDACTED\]/tv-fandoms/](https://www.pinterest.com/[REDACTED]/tv-fandoms/)
- Favorite Sports/Teams/Athletes
- Hobbies  
[https://www.pinterest.com/\[REDACTED\]/drawing/](https://www.pinterest.com/[REDACTED]/drawing/)  
[https://www.pinterest.com/\[REDACTED\]/baking/](https://www.pinterest.com/[REDACTED]/baking/)  
[https://www.pinterest.com/\[REDACTED\]/writing/](https://www.pinterest.com/[REDACTED]/writing/)  
[https://www.pinterest.com/\[REDACTED\]/acting/](https://www.pinterest.com/[REDACTED]/acting/)  
[https://www.pinterest.com/\[REDACTED\]/Cosplay ideas/](https://www.pinterest.com/[REDACTED]/Cosplay-ideas/)

### Other Info

- Spouse/Relatives  
[https://nuwber.com/person/563a\[REDACTED\]](https://nuwber.com/person/563a[REDACTED])
- Possible previous addresses (these are tied to spouse/relative might not apply to this individual)  
[https://nuwber.com/person/563a2\[REDACTED\]](https://nuwber.com/person/563a2[REDACTED])
- Mother's Maiden name  
[Intelius](#)  
[Whitepages](#)
- Phone Carrier  
[Whitepages](#)

ID: RP\_F

### Peripheral Layer

The peripheral layer (also called the identifier layer) is composed of biographic data such as name, age, and gender.

- Birthday  
<http://birth-records.mooseroots.com/d/b/>  
<https://www.facebook.com/profile.php>
- Name  
Gmail
- Email Address
- Other Phone#  
<http://69.175.16.130/~fair/wp-content/uploads/2016/02/>
- Home Address  
<http://www.peoplebyname.com/callerid/>  
<http://www.peoplebyname.com/people/>
- Photo  
<http://prwolfprints.com/features/2014/>  
<https://www.facebook.com/profile.php?id>

### Intermediate Layer

The intermediate layer (also called the demographic layer) is composed of affiliations such as religion, political views, and hometown.

- Hometown  
<http://www.whitepages.com/phone/1>
- Current City  
<http://www.whitepages.com/phone/1->
- Religion  
<https://www.facebook.com/profile.php?id=10>
- Employers
- Graduate school/college
- High School  
<http://prwolfprints.com/features/2014/>

### Core Layer

The core layer (also known as the favorites and tastes layer) is composed of emotions and values such as favorite music/books, sports, and activities.

- Favorite Music  
[Facebook Likes](#)
- Favorite Books
- Favorite Movies
- Favorite TV shows
- Favorite Sports/Teams/Athletes
- Hobbies  
<http://prwolfprints.com/features/2014/>

### Other Info

- Associates/Relatives  
<http://hodges-directory.us/directory.php?q=3nc9->  
<http://69.175.16.130/~fair/wp-content/uploads/>  
<http://www.peoplebyname.com/people/>
- Photo of home  
<http://www.remax.com/realestatehomesforsale/>
- Value of home  
<http://www.remax.com/realestatehomesforsale/>
- Birth State of parents  
<http://birth-records.mooseroots.com/d/b/>
- Mothers Maiden Name  
<http://birth-records.mooseroots.com/d/b/>
- Father Employment/Address History  
[https://pipl.com/search/?t=ODRhZGNmNmVjNTBjYzhhMmWQj](https://pipl.com/search/?t=ODRhZGNmNmVjNTBjYzhhMmWQj&in=8&q=815-444-8628&sloc=&l=&avatar=avatar-2) ZDA0  
&in=8&q=815-444-8628&sloc=&l=&avatar=avatar-2
- Languages  
[Facebook Likes](#)  
[Facebook Groups](#)
- Phone Carrier  
[Whitenpages](#)

ID: RP\_G

### Peripheral Layer

The peripheral layer (also called the identifier layer) is composed of biographic data such as name, age, and gender.

- Birthday
- Name
  - Possible full name? Looks like Laurel might be a middle name
  - [https://registrar.byu.edu/registrar/graduation/commencement/2015Su\\_Program.pdf](https://registrar.byu.edu/registrar/graduation/commencement/2015Su_Program.pdf)
  - [https://pipl.com/search/?q= \[redacted\] t&l=&sloc=&in=6](https://pipl.com/search/?q= [redacted] t&l=&sloc=&in=6)
- Found middle name
- Email Address
  - [https://pipl.com/search/?q= \[redacted\] t&l=&sloc=&in=6](https://pipl.com/search/?q= [redacted] t&l=&sloc=&in=6)
- Other Phone#
  - [https://www.truepeoplesearch.com/results?name= \[redacted\] &rid=0x0](https://www.truepeoplesearch.com/results?name= [redacted] &rid=0x0)
- Home Address
  - [http://www.homefacts.com/address/Illinois/Mchenry-County/Crystal-Lake/60014/\[redacted\]](http://www.homefacts.com/address/Illinois/Mchenry-County/Crystal-Lake/60014/[redacted])
  - [http://\[redacted\]](http://[redacted])
- Photo
  - [https://pipl.com/search/?q= \[redacted\] &l=&sloc=&in=6](https://pipl.com/search/?q= [redacted] &l=&sloc=&in=6)
  - [https://www.facebook.com/\[redacted\]](https://www.facebook.com/[redacted])

### Intermediate Layer

The intermediate layer (also called the demographic layer) is composed of affiliations such as religion, political views, and hometown.

- Hometown
  - [https://www.facebook.com/\[redacted\]](https://www.facebook.com/[redacted])
  - <http://www.switchboard.com/phone/1-508-268-8882>
- Current City
  - [http://www.homefacts.com/address/Illinois/Mchenry-County/Crystal-Lake/60014/\[redacted\]](http://www.homefacts.com/address/Illinois/Mchenry-County/Crystal-Lake/60014/[redacted])
  - [http://\[redacted\]](http://[redacted])
  - Intelius
  - [Maltego] [https://plus.google.com/116799-\[redacted\]](https://plus.google.com/116799-[redacted])
- Religion
  - [https://www.facebook.com/\[redacted\] about?section=contact-info&pnref=about](https://www.facebook.com/[redacted] about?section=contact-info&pnref=about)
  - [https://www.pandora.com/profile/thumbs/\[redacted\]](https://www.pandora.com/profile/thumbs/[redacted])
- Employers
- Graduate school/college
  - [https://registrar.byu.edu/registrar/graduation/commencement/2015Su\\_Program.pdf](https://registrar.byu.edu/registrar/graduation/commencement/2015Su_Program.pdf)
  - [https://scontent-sjc2-1.xx.fbcdn.net/v/t1.0-9/425672\\_484335884930102\\_1199939819\\_n.jpg?oh=d89-\[redacted\].1b1&oe=58963853](https://scontent-sjc2-1.xx.fbcdn.net/v/t1.0-9/425672_484335884930102_1199939819_n.jpg?oh=d89-[redacted].1b1&oe=58963853)
  - [Maltego] [https://plus.google.com/\[redacted\]](https://plus.google.com/[redacted])
- High School

### Core Layer

The core layer (also known as the favorites and tastes layer) is composed of emotions and values such as favorite music/books, sports, and activities.

- Favorite Music  
[Maltego] <https://www.pandora.com/profile/>  
<https://www.pandora.com/profile/thumbs/>  
<https://www.spokeo.com/social/profile?q=sw-xZ2AydC>
- Favorite Books  
<https://www.facebook.com/>
- Favorite Movies  
<https://www.facebook.com/>
- Favorite TV shows  
<https://www.facebook.com/>
- Favorite Sports/Teams/Athletes
- Hobbies  
[https://registrar.byu.edu/registrar/graduation/commencement/2015Su\\_Program.pdf](https://registrar.byu.edu/registrar/graduation/commencement/2015Su_Program.pdf)  
<https://www.pinterest.com/>  
Possible pinterest page  
<https://www.facebook.com/>

### Other Info

- Relatives/Associates  
<http://www.homefacts.com/address/Illinois/Mchenry-County/Crystal-Lake/60014/>  
<https://www.truepeoplesearch.com/results?name=&rid=0x0>  
[Maltego] <https://plus.google.com/116>
- Home Value  
<http://www.homefacts.com/address/Illinois/Mchenry-County/Crystal-Lake/60014/>
- Sale information on their house  
<https://www.co.mchenry.il.us/home/showdocument?id=>
- Wedding month/year  
[https://scontent-sjc2-1.xx.fbcdn.net/v/t1.0-9/549164\\_550420261654997\\_1441602441\\_n.jpg?oh=97160ada45c666e1](https://scontent-sjc2-1.xx.fbcdn.net/v/t1.0-9/549164_550420261654997_1441602441_n.jpg?oh=97160ada45c666e1)
- Spouse  
[https://scontent-sjc2-1.xx.fbcdn.net/t31.0-8/10355535\\_78](https://scontent-sjc2-1.xx.fbcdn.net/t31.0-8/10355535_78)
- Political views  
<https://www.facebook.com/>



## ID:RP\_H

### Peripheral Layer

The peripheral layer (also called the identifier layer) is composed of biographic data such as name, age, and gender.

- Birthday  
[Pipl-Email] Age [https://pipl.com/search?q=\[redacted\]7%40att.net&l=&sloc=&in=5](https://pipl.com/search?q=[redacted]7%40att.net&l=&sloc=&in=5)  
[Pipl] Date [https://www.linkedin.com/in/\[redacted\]](https://www.linkedin.com/in/[redacted])  
[Intelius](#)
- Name  
[Pipl] [https://pipl.com/search?q=\[redacted\]7%40att.net&l=&sloc=&in=5](https://pipl.com/search?q=[redacted]7%40att.net&l=&sloc=&in=5)  
[Intelius](#) (using phone #)
- Email Address  
[Intelius](#)
- Other Phone #  
[Pipl] [https://pipl.com/search?q=\[redacted\]7%40att.net&l=&sloc=&in=5](https://pipl.com/search?q=[redacted]7%40att.net&l=&sloc=&in=5)  
[Intelius](#)
- Home Address  
[Pipl] [https://pipl.com/search?q=\[redacted\]7%40att.net&l=&sloc=&in=5](https://pipl.com/search?q=[redacted]7%40att.net&l=&sloc=&in=5)  
[Intelius](#)
- Photo  
[Pipl] [https://pipl.com/search?q=\[redacted\]7%40att.net&l=&sloc=&in=5](https://pipl.com/search?q=[redacted]7%40att.net&l=&sloc=&in=5)  
[Pipl] [https://www.linkedin.com/in/\[redacted\]](https://www.linkedin.com/in/[redacted])  
[Pipl] [https://www.facebook.com/\[redacted\]](https://www.facebook.com/[redacted])

### Intermediate Layer

The intermediate layer (also called the demographic layer) is composed of affiliations such as religion, political views, and hometown.

- Hometown
- Current City  
[Pipl] [https://pipl.com/search?q=\[redacted\]7%40att.net&l=&sloc=&in=5](https://pipl.com/search?q=[redacted]7%40att.net&l=&sloc=&in=5)
- Religion  
[Pipl] [https://www.linkedin.com/in/\[redacted\]](https://www.linkedin.com/in/[redacted])  
[Pipl] [https://www.facebook.com/\[redacted\]](https://www.facebook.com/[redacted])
- Employers  
[Maltego] [https://twitter.com/\[redacted\]](https://twitter.com/[redacted])  
[Intelius](#)
- Graduate school/college  
[Pipl] [https://pipl.com/search?q=\[redacted\]7%40att.net&l=&sloc=&in=5](https://pipl.com/search?q=[redacted]7%40att.net&l=&sloc=&in=5)
- High School  
[Pipl] [https://pipl.com/search?q=\[redacted\]7%40att.net&l=&sloc=&in=5](https://pipl.com/search?q=[redacted]7%40att.net&l=&sloc=&in=5)  
[Intelius](#)

### Core Layer

The core layer (also known as the favorites and tastes layer) is composed of emotions and values such as favorite music/books, sports, and activities.


- Favorite Music  
[Maltego] <http://www.pandora.com/profile/stations/>  
[Pipl] <https://www.amazon.com/wishlist/>
- Favorite Books
- Favorite Movies
- Favorite TV shows  
[Maltego] <https://twitter.com/> /following
- Favorite Sports/Teams/Athletes
- Hobbies  
[Maltego] <https://klout.com/>  
[Maltego] <https://www.pinterest.com/> /following/  
[Pipl] <https://www.linkedin.com/in/>

### Other

- Associates  
[Maltego] <https://plus.google.com/1142/>  
[Pipl] <https://pipl.com/search/?q=>  
[Intelius](#)
- Username  
[Pipl] <https://pipl.com/search/?q=>
- Languages  
[Pipl] <https://www.linkedin.com/in/>
- Political Views  
[Pipl] <https://www.facebook.com/>
- Phone Carrier  
[Whitenpages](#)

**APPENDIX B. IRB DOCUMENTS**

One (1) printed unstapled copy must be submitted Page 1 of 15



**Application for the  
Use of Human Subjects**

Institutional Review Board for Human Subjects

BYU IRB USE ONLY:

ID #

Category

Type of Study

**Part A: Application Information**

**1. Title of Study**

Title

**2. Principal Investigator**

Name  Title

Phone  Email

Address

**3. Research Personnel** NOTE: Please include Principal Investigator in this section

Name	Department	Role	Study Responsibilities	CITI Login Name
Whitney Maxwell	Information Technology	Principal Investigator	Collect and analyze information of research subjects	wnielse
Chia-Chi Teng	Information Technology	Co-Investigator	Faculty member to advise in analysis and research	ccteng

The principal investigator and all personnel must have completed CITI training within the last 5 years prior to submission.

**4. Qualifications of Research Personnel**

Please provide a description of the training and experiences of the principal investigator and each of the Research personnel as it relates to the proposed research on the Personnel Biosketch forms.

- i. The Personnel Biosketch Forms should be attached as part G
- ii. Biosketches should be no longer than 2 pages
- iii. Note that curriculum vitae will not be accepted

**5. Funding Source**

(Note: If this research is associated with an active or pending research grant, attach a copy of the methodology section of the grant with this application.)

Application for Use of Human Subjects Ver. 06/14

a. How will this research be funded?

If other, please explain:

If the research is funded by an external sponsor, please complete the following information:

b. Name of Funding Source:

c. Status:  Funded  Pending

d. If funded, grant ID#:

e. Grant/Contract Title:

f. Principal Investigator listed on Grant/Contract:

**6. Research Duration** (Please note that the start date cannot occur before IRB approval)

Estimated start date (choose a date):

Estimated completion date (choose a date):

**Part B: Research Overview**

**7. Brief Research Summary (please limit to one page)**

Please provide a brief (no more than one page) summary for a general lay audience that includes background, rationale, and hypothesis for the proposed research.

This a Stage 1 Exploratory study which purpose is to determine the quantity and quality of information that can be uncovered through a combination of four common identifiers: email address, phone number, birth date, and name. The name and birth date identifiers primarily benefit the researcher, so that there is some assurance that information is being gathered about the right person. The email address and phone number identifiers are chosen because they are the ones that are often disclosed to third parties such as web sites and companies and due to their use, will reveal the majority of information being sought for this research. Prior research on information disclosure via identifiers such as emails and phone numbers indicate that these identifiers will be more than enough to draw the conclusions needed in this research design.

The research will be employing various Internet searches and research techniques to use those identifiers to uncover additional information on-line. Personal information available on-line is what is known as a "digital footprint." Though a digital footprint technically encompasses all personal information available on-line, this research will scope to only search for predetermined categories that vary in sensitivity. The categories are: additional phone numbers (cell,home,work), home address, photo, hometown, current city, religion, employers, graduate school/college, high school, favorite music, favorite books, favorite movies, favorite tv shows, favorite sports/teams/athletes, and hobbies.

The analysis will encompass analyzing the ease or difficulty in uncovering these categories as well as the value of each category. Given the sensitivity of this research, all data will be encrypted, secured and then destroyed at the end of the research period. The following research questions will be answered from the research:

1. How much information can be uncovered about a person given 1 or 2 details?

2. How can a digital footprint be captured and measured in size and impact using various metrics?  
3. How can a person best redact or further protect personal information from online sources?

**8. Research Categorization**

- a. Will the proposed research be conducted in a typical educational setting, involving normal educational practices, such as:  
- research on regular and special education instructional strategies, or  
- research on the effectiveness of or the comparison among instructional techniques, curricula, or classroom management methods

Yes  No

If yes, please list the activities below

- b. Are there activities (questionnaires, procedures, etc.) that subjects will perform regardless of their enrollment as a subject in the proposed research?

Yes  No

If yes, please list the activities below

- c. Will the proposed research gather data using educational tests, survey procedures, interview procedures, or observation of public behavior in such a way that investigators will be able to identify subjects? (Either directly or indirectly)

Yes  No

If yes, please list the activities below

- d. Are the subjects elected or candidates for public office?

Yes  No

If yes, please explain

- e. Will the research involve the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens that are publicly available or where the information is recorded by the investigator in such a manner that subjects cannot be identified by the investigators?

Yes  No

If yes, please explain

**9. Location of Research**

Please indicate all locations where the proposed research will be conducted (i.e., room numbers, local address, institution, city, state, country, etc.)

Online only

**10. Research Subject Information**

- a. Number of Subjects
- b. Gender of Subjects
- c. Ages of Subjects

d. Identify any of the following vulnerable populations that will be included in the proposed research.

<input type="checkbox"/> Children	<input type="checkbox"/> Institutionalized
<input type="checkbox"/> Pregnant Women	<input type="checkbox"/> Students Enrolled in Your Class(es)
<input type="checkbox"/> Prisoners	<input type="checkbox"/> Mentally Disabled Persons*
<input type="checkbox"/> Economically Disadvantaged Persons	<input type="checkbox"/> Educationally Disadvantaged Persons

*\*An individual who is unable to provide informed consent for themselves*

e. If the proposed research will include a vulnerable population, as identified above, please provide justification for their inclusion.

f. Please provide a description of the subject population. Specifically explain the criteria that will be used to include or exclude potential subjects from participating in the proposed research.

Subjects will need to represent a variety in age range. At least one subject will be chosen that falls into each of the following categories:  
 Age 18-24, 25-29, 30-40, 41+

g. Please provide an explanation and justification for the number of subjects (Part B, 10a) to be enrolled in the proposed research (e.g., power/sample size analysis, citation of comparable studies from the literature).

The number of subjects is determined by representing each of the age groups with 2-3 research subjects. The number will be capped at 15 since research involving each subject will be time and resource intensive. While this is a small sample, this research is a Stage I Exploratory project. It is an initial attempt at a proof-of-concept design that could be built on for future research and larger impact.

**11. Non-English Speaking Subjects**

In accordance with federal regulations, subjects cannot be excluded from research on the basis of race, sex, age, language, or disability status.

- a. Will the proposed research likely recruit and/or involve subjects or parents/guardians of subjects who are not fluent in English?

Yes  No

If yes, please indicate the languages:

If yes, please submit an English version of the consent form AND a version of the consent form translated into the appropriate language(s) as attached documents in part E.

- b. If you believe the proposed research requires the exclusion of a particular non-English speaking subject population please provide a justification and explanation below. (Note: investigator inconvenience and/or additional effort associated with producing bilingual documents will not be considered as an appropriate justification)

**12. Subject Compensation**

- a. Will subjects be compensated for participation?

Yes  No

If yes, please describe the form of compensation (i.e., cash, check, gift certificate, voucher, 1099, etc.).

- b. If compensation is monetary please provide the amount of compensation.

- c. Please provide a brief justification for the form and amount of compensation provided to subjects.

- d. Please describe when and how compensation will be provided to subjects.

- e. Will compensation be prorated?

Yes  No

If yes, please describe how proration will occur.

- f. For research involving students, will compensation be provided in the form of class credit to students that choose to participate?

Yes  No

If yes, describe how students who choose not to participate in the research may earn class credit. The alternative method to earn class credit should be comparable in time and effort required for research subjects to complete study tasks.

## Part C: Research Proposal

### 13. Research Question

What do you intend to measure, observe, describe, demonstrate, confirm, etc. from the proposed research?

The following research questions will be answered from the research which is a Stage 1 Exploratory project:

1. How much information can be uncovered about a person given a small number of details?
2. How can a digital footprint be captured and measured in size and impact using various metrics?
3. How can a person best redact or further protect personal information from online sources?

### 14. Hypothesis

If applicable, state your hypothesis for each of the research questions listed in Part C, 13

1. If a small number of details are known about a given subject, then they can be used to uncover a large amount of personal information
2. If a digital footprint is only collection of digital personal information then it should be able to be represented in a graphical format.
3. If there is personal information that is available online, then it should be able to be redacted at the person's request.

### 15. Background and Significance (please limit to 2 pages)

This section should provide clear, logical and sufficient support from the scientific literature that provides a rationale for the proposed aims and hypotheses stated above.

The following is a question that sparked this research: how big is my digital footprint? For someone as technically savvy as myself, I thought my response to this question would be simple. I of all people, should know how much of my information was out there. The answer I thought I had, was wrong. Oh so very wrong. Preliminary research on the topic was shocking. The book Data and Goliath quickly becomes a depressing read after just a few pages. Interestingly, one of the biggest violators of personal privacy are cell phones. For instance, the UK Company Cobham makes it possible to send a "blind" call to a phone. This call doesn't cause the phone to ring but it does force it to transmit on a specific frequency and thereby give up its location. As if knowing your



exact location isn't scary enough, in 2012 researchers were able to use cellular data to predict where people would be 24 hours later, accurate to within 20 meters. And location isn't the worst of it. 95% of Americans can be identified by name from just four time/date/location points (Schneier, 2015). One of the uses for this type of information is marketing. Companies want to target ads to people who are close in proximity to one of their stores. Targeting ads to people based on time and location along with their personal information has turned marketing into a whole new ball game, and a wealthy one at that.

As you can imagine, marketing isn't the only source that ingests metadata such as location. Governments use surveillance and metadata to track everything from potential terrorists to those who don't clean up after their dog. According to NSA general counsel Stewart Baker, "Metadata absolutely tells you everything about somebody's life. If you have enough metadata you don't really need content." Building on this the former NSA and CIA director Michael Hayden confirmed, "We kill people based on metadata" (Schneier, 2015). Maybe this isn't as chilling if you believe you are innocent. Surely the old adage, "If you did nothing wrong you have nothing to hide" implies that privacy is only valuable to wrongdoers. However, there are plenty of things we do each day that we want to keep private. And while you may not feel you are on the bad side of a government your race, gender, religion, or political affiliation might put you in the cross hairs of a hate crime. If the government can target potential terrorists based on metadata, just think of who might be targeting you.

While it might be easy to recommend swearing off technology forever and living in the bush in Africa most people still want to enjoy the benefits of technology and the internet. Unfortunately just using "incognito mode" on your browser won't do the trick. Cleaning up and minimizing a digital footprint isn't hopeless but it does take some work. In the book, How To Disappear (Ahearn & Horan, 2010), Ahearn enumerates various ways to eliminate digital footprints and even includes ways on how to set up false trails for anyone trying to look for you. There has been research done on how to cover or uncover personal information, but I haven't found any that try to actually quantify and measure its impact.

**16. Subject Recruitment and Informed Consent**

a. Which of the following tools will be used to recruit subjects? (Check all that apply)

<input type="checkbox"/> Posted/Distributed Flyer	<input type="checkbox"/> Third Party / Non-study Personnel / Professional recruiters
<input type="checkbox"/> Classroom Announcement	<input type="checkbox"/> Website / SONA / Social Media / Online
<input type="checkbox"/> Email to Subjects	<input checked="" type="checkbox"/> Other: Convenience Sampling

b. Please describe the initial contact with potential subjects and the process to obtain their consent.

All subjects are given by personal referrals from friends and family. Once I have their name and contact information I will contact them to further explain my research and after they have confirmed their interest I will send them the official consent form to officially participate in the research study. The script for the PI is as follows:

PI: Hello, this is Whitney Maxwell and I am contacting you about participating in research for my Master's Thesis. I understand you showed some interest in participating in my research regarding information privacy is that correct? <if yes, continue. if no, thank them for their initial interest and end conversation>

I wanted to start by going over the consent form that you will need to sign to officially participate in the study. It contains an introduction, procedures of the research and then risks and benefits of participating. <I will then read the consent form with them> Now that we have read the consent form, are there any other questions I could answer? If you think of any questions feel free to reach out to me and I will be happy to answer them.

c. Are you requesting a Waiver or Alteration of Informed Consent?

Yes  No

If yes, the Request for Waiver/Modification of Consent form (see ORCA website) must be completed and submitted in Part E with this application.

d. Have you attached the necessary consent forms in Part E of the application?

Yes  No

e. Have you attached a copy of all recruitment materials (i.e., scripts, flyers, screen shots of online material, etc.) in Part H of the application?

Yes  No

### 17. Research Methods

a. Write a complete description of all procedures involving human subjects in the proposed research. This description should encompass the experimental course of a subject from their entry into the study to their completion of the study.

If the subject agrees to participate in this **Stage 1 Exploratory** research study, the following will occur:

- The subject will provide the following information to the researcher: full name, phone number, birth date and email address. I have chosen these types of information since they are commonly found in public domains and are often disclosed to third parties. The researcher will use these provided identifiers to attempt to discover as many of the following categories as possible: additional phone numbers (cell,home,work), home address, photo, hometown, current city, religion, employers, graduate school/college, high school, favorite music, favorite books, favorite movies, favorite tv shows, favorite sports/teams/athletes, and hobbies. I will begin with the first piece of information (full name), and proceed to uncover as many of the previously listed categories as possible. If I discover a category such as home address, I will only mark "home address" and not the actual address in an effort to protect the original information. Once I have discovered as many of the predetermined categories as possible, I will then build on that data by adding the second piece of information (date of birth) and ideally uncover more categories. I will continue to progressively add the following information (phone number and email address) one at a time and document what categories are further uncovered with each iteration. Once all four pieces of information are in play I will try and exhaust all methods and tactics to uncover the remaining categories. Each person that I will profile will sign a letter of consent which will be approved by the BYU IRB, allowing me to use any legal method of obtaining information.

This research is designed to capture information that can be found using various Internet searches and websites that are available to everyone either for free or for a very low cost (low meaning less than \$5). Oftentimes the search services that require payment still acquire data from free sources but charge payment for the convenience of listing it all in one place. It is not designed to use highly invasive tactics or expensive services. The majority of web sites that I use were found through Google.

I will document all the tools and techniques used to uncover personal information and use it to develop a digital footprint metric. It will take into account the probability of uncovering each category, the number of categories uncovered, and the personal value of the categories (i.e. home address has higher personal value than favorite sport).

Lastly, I will document methods of information protection or removal and develop recommendations for redacting or reducing visibility of personal information.

- After the researcher has concluded their research, they will contact the subject via phone or email to review all personal information they uncovered during the research process and disclose how it was uncovered.

• Total time commitment will be about thirty (30) minutes

- b. Many studies use multiple instruments/questionnaires/surveys as part of the research methodology. If applicable to the proposed research, list each instrument/questionnaire/survey that will be administered to subjects and provide a rationale for the inclusion of each one.

- c. Please indicate how many times human subjects will perform research activities.

Two. There is the initial visit to go over the consent form and answer questions, and the follow up visit to go over the results.

- d. How much time will each visit take?

Approximately 15 minutes

- e. What is the total time required for a subject to complete the proposed research?

30

- f. Will the research include the use of existing data, research records, patient records, and/or human biological specimens?

Yes  No

If yes, please provide a description of the data and the source of the data. (If applicable, attach the data transfer agreement in part H)

As part of the participation agreement, the research subject will supply specific details such as email and phone number for initial contact and follow up.

If the data has identifiable subject information associated with it, please describe how the data will be de-identified.

Each piece of identifiable information will be changed or modified to eliminate any chance that it could be traced to the original subject. All names will be replaced with a unique numerical ID and any other information such as birth dates/addresses/etc. will be modified. Any information that cannot be modified such as photos will be given a generic name (such as ID\_photo\_1) in a list and will not be included in any written reports.

- g. Will the research involve deception or less than full disclosure?

Yes  No

If yes, please justify the need for deception in the proposed research.

h. Will the research require accessing student educational records?

Yes  No

If yes, please describe the procedures you will employ to access the information. Specify the information you will request.

i. Does the research involve audio/video recording or photography?

Yes  No

If yes, please describe how and in what situations the recordings/images will be used. If applicable, attach the video/photographic release form in Part H.

Photos will be obtained primarily through social media or other websites. They will be obtained as a proof of concept but they will not be published in any written reports.

j. Will subjects be followed after completion of the proposed research?

Yes  No

If yes, please explain.

k. Have you included a copy of all instruments, questionnaires, surveys, and/or interview questions to be used for this research in Part F of the application?

Yes  No

**18. Data Analysis**

Whether quantitative or qualitative, please provide a detailed description of the research design/statistical design and data analysis plan for the proposed research.

This research involves both quantitative and qualitative data analysis. Each piece of personal information uncovered during the research period will be marked with tags such as: discoverability, value and risk. All data will be compiled and analyzed on a per subject basis and then compiled and analyzed again across all subjects. Each of the tags associated with each piece of information will have an effect in the graphical and statistical representation of the digital footprint.

**19. References**

Please provide pertinent references from the literature that support the rationale and hypothesis for the proposed research. (Limit to one page.)

Ahearn, F. M., & Horan, E. C. (2010). How to disappear : erase your digital footprint, leave false trails, and vanish without a trace. Guilford, Conn.: Lyons Press.

Bansal, G., & Zahedi, F. M. (2015). Trust violation and repair: The information privacy perspective. *Decision Support Systems*, 71, 62—77. <http://doi.org/10.1016/j.dss.2015.01.009>

Conti, G. (n.d.). *Googling security : how much does Google know about you?* Upper Saddle River, NJ : AddisonWesley, 0009.

Digital Footprints: An Internet Society Reference Framework I Internet Society. (n.d.). Retrieved December 4, 2015, from <http://www.intemetsociety.org/doc/digital-footprints-internet-society-reference-framework>

Hewson, K. (2013). What size is your digital footprint? *Phi Delta Kappan*, 94(7), 14—15. Retrieved from <https://www.lib.byu.edu/cgi-bin/remotefauth.pl?url=http://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=86873369&site=ehost-live&scope=site>

Schneier, B. (2015). *Data and Goliath : the hidden battles to collect your data and control your world.*

Teurlings, J., & Stauff, M. (2014). Introduction: The Transparency Issue. *Cultural Studies ++ Critical Methodologies*, 14(0), 3-10. <http://doi.org/10.1177/1532708613519184>

## 20. Benefit to Subjects and Society

- a. Please describe any benefits that research subjects will receive as a direct result of their participation in the proposed research. (Note: compensation is not considered a benefit)

The research subjects will benefit by gaining a realistic understanding of their current level of personal security by seeing a snapshot of what personal information is available on-line In addition, they will be provided with best practices of how to better secure personal information and thereby improve their level of personal security.

- b. Please describe how this research may provide benefit to scientific knowledge, a specific discipline, and/or society in general.

While there is existing research on digital footprints, there haven't been any advances in taking that data and formatting it into statistical or graphical formats which are easier for people to understand. Most people have a level of awareness when it comes to the fact that some of their information is publicly available, however, understanding the full scope of where their information is and the risks associated with it is more difficult to grasp. This research not only will advance the field of Information Technology in the cybersecurity sphere but also the general population as well.

## 21. Risks

Generally, risk assessment in research considers the harm, trauma, discomfort, stress, or any other undesirable or untoward consequence of being a research subject whether anticipated or unexpected. Risk may take the following forms: physical, psychological, emotional, financial, and/or social. This list is representative but not comprehensive.

Please consider the following questions based on the experiences that subjects might encounter through participation in the proposed research.

- a. Whether great or small, please describe the potential risks to subjects that participate in the proposed research.

Information about the research subjects will be gathered and analyzed using targeted Internet searches and on-line research tools. Personal information about the subjects that could be uncovered includes but is not limited to: additional phone numbers (cell,home,work), home address, photo, hometown, current city, religion, employers, graduate school/college, high school, favorite music, favorite books, favorite movies, favorite TV shows, favorite sports/teams/athletes, and hobbies. The details of this information will be seen by the researcher or co-researcher only but will not be included in any research notes or publications.

Should any criminal activity be discovered during the research it will be reported to city authorities including but not limited to: fraud, illicit drug use, and sexual exploitation of children. For a more comprehensive list please see <https://www.justice.gov/usam/organization-and-functions-manual-10-partial-list-federal-matters-investigated-fbi>.

At the conclusions of the study the researcher will provide methods and guidelines to the research subject of how to further protect and lock down their personal information. The researcher will also replace personally identifiable information with generic IDs (e.g. Subject A) in all collected information and not retain any information that could be tied back to the research subject directly.

- b. What steps will be taken by investigators to reduce the aforementioned risks?

As mentioned previously, no details of personal information aside from the initial 4 identifiers will be recorded. Only the name of the category of information will be noted and the details of how that category of information will be obtained.

When used in analysis your name will be replaced by a unique ID and all personally identifiable information will be removed in written reports. Once all personally identifiable information has been used for analysis it will be destroyed.

The research done on the subject is not designed to be comprehensive or invasive. The researcher will begin with a set of web sites to identify as many categories as possible. They will be limited to what they can find with Internet searches and they will not be actively trying to breach personal boundaries or manipulate any companies/friends/acquaintances in an attempt to acquire more information, such as trying to "friend" them on social media sites or making phone calls. The researcher will encourage the research subjects to consider the risks carefully and withdraw their consent if they feel uncomfortable with information that could be discoverable online.

## 22. Confidentiality

The following questions address your efforts to maintain subject confidentiality in the proposed research. Protecting hard copy data may involve de-identification of data and secured storage locations and conditions. Protecting electronic data may involve a secure network, password access, and data de-identification/encryption.

- a. Describe what type of hard copy data will be generated by the proposed research (i.e., notes, audio/video tapes, questionnaires, etc.).

Written reports including a written thesis will be generated from the proposed research.

- i. Where will this hard copy data be stored and how will it be protected?

The data contained in the hard copies will be aggregated and any personally identifiable information will be removed or replaced by a generic ID. No sensitive data will be contained in a hard copy format.

- b. Describe what type of electronic data will be generated by the proposed research (i.e., computer files/spreadsheets, questionnaires, images, video, audio/mp3 files, etc.).

The research will involve written notes of Internet search results, and categories collected.

- i. Where will this electronic data be stored and how will it be protected?

All digital copies of research will be stored in OneNote which is protected by user name and password as well as requiring smart card authentication. The hard drive containing the data will encrypted by BitLocker.

- c. How long will data from this study be maintained by the principal investigator?

The data will be collected for analysis and retained for 3 years according to policy.

- d. Will raw data be made available to anyone other than the principal investigator and the immediate study personnel?

Yes  No

If yes, describe the rationale and procedure for sharing study data. Include a description of what will be shared and with whom. Specify the protections in place to transfer the data.

**Part D: Researcher Agreement**

The research study involves the use of human subjects. I understand the university's policy concerning research involving human subjects and by submitting this application I agree to:

- Obtain voluntary and informed consent of subjects who are to participate in this project.
- Report to the IRB any unanticipated effects on subjects which become apparent during the course of, or as result of, the experimentation and the actions taken.
- Cooperate with members of the committee charged with continuing review of this project.
- Obtain prior approval from the committee before amending or altering the scope of the project or implementing changes in the approved consent document.
- Maintain the documentation of consent forms and progress reports as required by institutional policy for three years.
- Safeguard the confidentiality of research subjects and the data collected when the approved level of research requires it.

Signature of the Principal Investigator Whitney N Maxwell Date

Note: Submissions from student principal investigators require the signature of their faculty sponsor (see below).

I have read and reviewed this proposal and certify that it is ready for review by the IRB. I have worked with the student to prepare this research protocol. I agree to mentor the student during the research project.

Printed Name of Faculty Sponsor

Signature of Faculty Sponsor \_\_\_\_\_ Date

**IRB Forms Submission Instructions:**

1. Check this form for **completeness**, detail, and accuracy.
2. Save this form for your records.
3. Applications received **on or before the 20th** of each month will be reviewed by the committee on the first Thursday of the following month.
4. Upon receipt of your proposal by the IRB office, you will be assigned a proposal ID number that should be used in all future correspondence concerning your proposal.



**Additional Parts**

Please attach the appropriate Parts as described below.

Applications that do not contain all of Parts A-D and all necessary Additional Parts will be returned to the applicant without a review.

Forms are available at <http://www.orca.byu.edu/irb/BeginApplication.php>.

**Part E**

- Consent documents
- Assent documents
- Parental permission documents
- Request for Waiver or Modification of Consent form

**Part F**

- All questionnaires, surveys, interview questions, discussion questions

**Part G**

- Biosketches of all Research Personnel (each Biosketch should be no longer than 2 pages)

**Part H**

- Letters of support from sponsoring institutions/organizations, if applicable
- Photographic Release
- Video Release
- Recruiting Materials (including scripts, flyers/posters, letters, screen shots of online recruiting materials, etc.)
- A copy of research grant methodology section, if applicable

# Brigham Young University

## Consent to be a Research Subject

### Introduction

---

This research study is being conducted by graduate student Whitney Maxwell under the direction of IT faculty member Chia-Chi Teng at Brigham Young University. The purpose of this study is to determine how much information can be gathered for a specific person through online research in order to analyze what is known as a “digital footprint” i.e. personal details available online. The research will be scoped to only search for the following categories of information: phone numbers (cell,home,work), home address, photo, hometown, current city, religion, employers, graduate school/college, high school, favorite music, favorite books, favorite movies, favorite tv shows, favorite sports/teams/athletes, and hobbies. While the intent is to only discover the aforementioned categories, it is possible that there will be other information that is disclosed in the process. Any additional categories will be noted.

The analysis will encompass the probability in uncovering these categories, the number of categories discovered, and the sensitivity of each category. Given the sensitivity of this research, all data will be encrypted, secured and then destroyed at the end of the research period. Details are outlined in the Confidentiality section.

You were invited to participate because of your interest in the topic and your age demographic.

### Procedures

---

If you agree to participate in this research study, the following will occur:

- You will provide the following identifying information to the researcher: full name, phone number, birthdate and email address. The researcher will then employ Internet searches and online tools to use those identifiers to uncover as many of the predefined categories as possible. Please note that when the researcher discovers a category such as Employers during their research, that they will only note the name of the category “Employers” in their notes and not the actual information to retain privacy.
- After I have concluded the research, I will contact you via phone to review all personal information they uncovered during the research process and disclose how it was uncovered.
- Total time commitment will be about thirty (30) minutes

### Risks/Discomforts

---

Information about you will be gathered and analyzed using on-line searches and tools. Personal information that could be uncovered includes but is not limited to: phone numbers (cell,home,work), home address, photo, hometown, current city, religion, employers, graduate school/college, high school, favorite music, favorite books, favorite movies, favorite tv shows, favorite sports/teams/athletes, and hobbies. Please note that the details of this information will not be recorded or published.

Ver. 12/12

While the intent is not to look for any illicit or criminal activity, it is possible that incriminating evidence could be discovered in the research process. Should any illegal documents be discovered during the research (such as child pornography) you will be reported immediately to the FBI.

At the conclusions of the study the researcher will provide methods and guidelines of how to further protect and lock down personal information. The researcher will also scrub all collected information and not retain any information that could be tied back to you directly.

## Benefits

---

You will benefit by gaining a realistic understanding of your current level of personal security by seeing a snapshot of what types of information is currently available on-line. In addition, you will provided with best practices of how to better secure personal information and thereby improve your level of personal security.

## Confidentiality

---

Given that all information gathered as part of this research study will be personal , all data will be generalized where possible and stored on a computer with an encrypted hard drive and multi-factor authentication.

When used in analysis your name will be replaced by a unique ID and all personally identifiable information will be scrubbed in written reports. Once all personally identifiable information has been used for analysis it will be destroyed.

## Participation

---

Participation in this research study is voluntary. You have the right to withdraw at any time or refuse to participate entirely.

## Questions about the Research

---

If you have questions regarding this study, you may contact Whitney Maxwell at whitneymaxwell@gmail.com for further information or Information Technology faculty advisor Chia-Chi Teng at ccteng@byu.edu.

## Questions about Your Rights as Research Participants

---

If you have questions regarding your rights as a research participant contact IRB Administrator at (801) 422-1461; A-285 ASB, Brigham Young University, Provo, UT 84602; irb@byu.edu.

## Statement of Consent

---

I have read, understood, and received a copy of the above consent and desire of my own free will to participate in this study.

Name (Printed): \_\_\_\_\_ Signature: \_\_\_\_\_ Date: \_\_\_\_\_

Ver. 12/12