The University of Texas at Austin
Center for Identity

Is Your Phone You? How Privacy Policies of Mobile Apps Allow the Use of Your Personally Identifiable Information

Kai Chih Chang
Razieh Nokhbeh Zaeem
K. Suzanne Barber

March 2020

UTCID Report #2002

Is Your Phone You?

How Privacy Policies of Mobile Apps Allow the Use of Your Personally Identifiable Information

KAI CHIH CHANG, The University of Texas at Austin
RAZIEH NOKHBEH ZAEEM, The University of Texas at Austin
K. SUZANNE BARBER, The University of Texas at Austin

People continue to store their sensitive information in their smart-phone applications, knowingly or more often unknowingly. Users seldom read an app's privacy policy to see how their information is being collected, used, and shared. In this paper, using a reference list of over 600 Personally Identifiable Information (PII) attributes, we investigate the privacy policies of 100 popular health and fitness mobile applications in both Android and iOS app markets to find the set of personal information these apps collect, use and share. The reference list of PII was independently built from a longitudinal study at The University of Texas investigating thousands of identity theft and fraud cases where PII attributes and associated value and risks were empirically quantified. This research leverages the reference PII list to identify and analyze the value of personal information collected by the mobile apps and the risk of disclosing this information. We found that the set of PII collected by these mobile apps covers 35% of the entire reference set of PII and, due to dependencies between PII attributes, these mobile apps have a likelihood of indirectly impacting 70% of the reference PII if breached. For a specific app, we discovered the monetary loss could reach $1M if the set of sensitive data it collects is breached. We finally utilize Bayesian inference to measure risks of a set of PII gathered by apps: the probability that fraudsters can discover, impersonate and cause harm to the user by misusing only the PII the mobile apps collected.

Additional Key Words and Phrases: Identity, Internet of Things, Privacy Policy, Mobile Apps

ACM Reference Format:
Kai Chih Chang, Razieh Nokhbeh Zaeem, and K. Suzanne Barber. 2020. Is Your Phone You?: How Privacy Policies of Mobile Apps Allow the Use of Your Personally Identifiable Information. 1, 1 (January 2020), 15 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In this era of advanced Internet of Things (IoT), smart-phones are rapidly becoming the mobile platform of choice for users around the world. There are many factors behind this growth, but one of the main reasons is the large number of mobile phone applications in the market. By 2018, more than 4 billion applications were available for download on the Android Google Play and iOS App Store [12]. An American checks his or her mobile phone every 12 minutes on average [25]. We know that smart-phones, these IoT devices, are inseparable from our lives. With the rise of concerns of health and workout in recent years, health care and fitness devices have been widely used. Among those devices, mobile phone is the most common device that a person can obtain. Bunchs of health care and fitness applications have been widely downloaded. The widespread use of these apps, however, poses a concerning challenge to users' privacy.

Authors' addresses: Kai Chih Chang, The University of Texas at Austin, Austin, Texas, kaichih@identity.utexas.edu; Razieh Nokhbeh Zaeem, The University of Texas at Austin, Austin, Texas, razieh@identity.utexas.edu; K. Suzanne Barber, The University of Texas at Austin, Austin, Texas, sbarber@identity.utexas.edu.

XXXX-XXXX/2020/1-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Personally Identifiable Information (PII) is commonly used in both physical and cyber worlds to perform personal authentication. Through various mobile applications, people store sensitive information in applications or upload them to the Internet. On Facebook only, 400 new users join and 147,000 photos are uploaded every minute [5]. The information entered during registration into the Facebook app and the information contained in the photos were all actively and voluntarily given out by users, but mobile applications also collect passive information such as location, IP address, and web browsing history.

Some applications collect personal data without a privacy policy that advises how the data is collected, shared, used, and safeguarded. Other applications follow industry-accepted privacy policies, but sometimes use shadow ad tracking. Many apps also send information to third parties, including Google and mobile marketers. The problem with these behaviors is that (1) they are widespread, (2) users are not told that the app is tracking them, and (3) they have very little control over the information collection and use process [6]. For example, on average, a health and fitness application pings 5.6 third-party trackers in the first minute of its use [24]. Over 30% of such health apps have no privacy policy [8]. The current privacy controls are no better than a game of Whac-A-Mole. Just as one mole is knocked down, there are many more moles coming out, spying on users' private information through apps. Before we can protect users' private information, we first need to know where this information is collected and how it is shared or used.

Google Play requires that all Android apps that collect and handle personal or sensitive user data have a Privacy Policy in place. For iOS devices, as of October 3, 2018, App Store Connect requires all new apps and app updates to have a privacy policy, submitted as part of the App review process for submission to the App Store. In addition, links or text for the App Privacy Policy can only be edited when a new version of the app is submitted.

In this paper, we examine 100 popular health and fitness apps in two widely-used smart-phone platforms—Android and iOS—to find out which PII attributes are collected by these applications. We then compare the list of PII attributes with an independently built comprehensive list of 627 PII gathered through manual investigation of over 6,000 identity theft news reports. Furthermore, we leverage our previous work, the UT Center for Identity (CID) Identity Ecosystem, to gain a deeper comprehension of the PII that apps collect and the relationship between various PII attributes. The UT CID Identity Ecosystem developed at the Center for Identity (CID) at the University of Texas at Austin has constructed a graph-based model of people, devices, and organizations. It provides a framework for understanding the value, risk and mutual relationships of PII attributes. Every attribute is modeled as a graph node which has several properties, while the relationships between PII attributes are modeled as edges. The UT CID Identity Ecosystem is capable of answering several queries about PII through Bayesian inference. By querying UT CID Identity Ecosystem, we produce interesting insights about what would happen if the information gathered by a smart-phone is breached. Particularly, we infer the probability that a frustrater can impersonate the mobile user by using this information for authentication purposes.

This paper makes the following contributions:

(1) This is also the first work to use an independently built, comprehensive list of PII attributes to evaluate the privacy of mobile apps.
(2) Having access to UT CID probabilistic models and Bayesian inference tool, this is the first work to take advantage of Bayesian inference to answer fundamental queries about app privacy.

The remainder of this article is structured as follows. Section 2 presents the related work of privacy in the Internet of Things (IoT). Section 3 introduces our previous work which motivates this paper. Section 4 provides a brief description of our experiment setup and methodology. Section

5 includes a comprehensive analysis for evaluation. Section 6 introduces the UT CID Identity Ecosystem and how we utilize it to help with our analysis. Section 7 concludes our research and gives directions for future work.

2 RELATED WORK

Related work on mobile devices and PII privacy can loosely be divided into the following categories: techniques about *accessing* sensitive data, techniques about detecting PII *transmission*, and techniques about reducing or avoiding digital privacy risk.

First, a great body of research is dedicated to the possibility of mobile apps accessing sensitive data. D'Orazio et al. [9] proposed a generic process for iOS medical applications to identify their vulnerabilities and design weaknesses by decrypting the executable file for static analysis and dynamic analysis. Zhao et al. [29] modeled the location probing attacks and proposed some approaches to perform such large-scale attacks on Location-Based Social Network (LBSN) applications. They applied these approaches on eight LBSN applications and showed that they were able to collect users' location information, which could be exploited to invade users' privacy. Ackerman [2] analyzed 23 free and 20 paid apps across categories like behavioral health, health and fitness, diet, pregnancy, quiting smoking, etc. They analyzed and measured mobile health and fitness applications by objective and subjective criteria. Dehling et al. [8] surveyed most popular mHealth apps in the Apple iTunes Store and the Goolge Play Store to assess the availability, scope, and transparency of mHealth app privacy policies and found out that of the 600 most commonly used apps, only 183 had privacy policies. Rowan et al. [22] provided the results of a privacy policy comparison including application permissions and several metrics used to assess the current state of privacy policies in the health and fitness mobile application market. Liu et al. [18] examined web sites of the Fortune 500 and showed that only slightly more than 50 percent of Fortune 500 web sites provide privacy policies on their home pages. Harris [15] issued recommendations for mobile application developers and the mobile industry to safeguard consumer privacy. This work provided guidance on developing strong privacy practices, translating these practices into mobile-friendly policies, and coordinating with mobile industry actors to promote comprehensive transparency. The above work discussed how mobile apps could access sensitive data but did not explicitly show the exact set of PII collected by mobile applications.

Second, researchers have also studied the transmission of personal data from mobile devices to any third parties. The Wall Street Journal conducted a study on 101 popular smart-phone applications in 2010 and analyzed their data traffic through Wi-Fi connection and showed that 56 transmitted the phone's unique device ID to other companies without users' awareness or consent. In addition, 47 apps transmitted the phone's location in some way and 5 sent age, gender and other personal details to outsiders (e.g., ads company) without informing users [26]. Smith [23] studied a number of iPhone apps from the "Most Popular" and "Top Free" categories in Apple's App Store. For these applications, he collected and analyzed the data being transmitted between installed applications and remote servers using several open source tools and found that 68% of these applications were transmitting unique device identifiers (UDIDs) to servers under the application vendor's control each time the application is launched. Egele et al. [10] constructed a control flow graph of downloaded and decrypted iOS apps, to determine if there is a path from "sources" of sensitive data to "sinks" where the information can leave the device. Their analysis on 1400 iOS apps detected hundreds of apps accessing and sending the UDID. Papageorgiou et al. [20] investigated user's privacy exposure in m-health apps and have studied the spread of users' personal data and, mainly, the final parties receiving these data. They performed both static and dynamic analysis on 20 popular m-health apps and found out that the majority of the analyzed apps did not meet the expected standards for security and privacy, thus endangering their users' sensitive personal data. Anthi et al. [4] examined 96 free mobile applications across 10 categories, in both the Apple App Store and Google Play Store, to investigate how securely they transmit and handle user data. They performed wireless packet sniffing and a series of man-in-the-middle (MITM) attacks to capture PII. During the MITM attacks, they used a variety of methods to try to decrypt the transmitted information. According to their study, 60% of iOS and 25% of Android applications transmit unencrypted user data over the Wi-Fi network and they observed that data is being forwarded to third party domains. Han et al. [14] proposed a study of real-world tracking via mobile apps in which they measured how 20 participants were tracked over three weeks as they used their Android smart-phone apps. They instrumented the phones with dynamic taint tracking to record communications that exposed identifying information, and inspected web cookie databases. They found that 36% of the sites (655 out of 1824) that their study apps programmed to contact are tracking users. Enck et al. [11] proposed modifying the Android OS such that taint values can be assigned to sensitive data and their flow can be continuously tracked through each app execution, raising alerts when they flow to the network interface. They imposed a runtime overhead because it runs continuously for all applications and hence the authors tested it on a set of thirty popular Android apps reporting that many of them leak privacy sensitive data. Above studies showed the interest of third parties in those PII but they did not talk about the reason that third parties crave those personal information. Are those PII valuable or strong authenticators? Moreover, they did not mention the outcome of letting PII being exploited by third parties.

Third, to protect user privacy, researchers have begun to explore techniques and analysis for mitigating digital privacy risk. Agarwal et al. [3] proposed ProtectMyPrivacy (PMP), a crowdsourced recommendation engine, to analyze manual protection decisions, and use them to provide iOS app privacy recommendations. They showed that based on the protection decisions contributed by their users they can recommend protection settings for over 97.1% of the 10,000 most popular apps. Hornyack et al. [16] introduced AppFence. They implemented data shadowing, to prevent applications from accessing sensitive information that is not required to provide user-desired functionality, and exfiltration blocking, to block outgoing communications tainted by sensitive data. Rao et al. [21] presented Meddle, a platform that leverages virtual private networks (VPNs) and software middleboxes to improve transparency and control for Internet traffic from mobile systems. By controlling privacy leaks and detecting ISP interference with Internet traffic they found PII leaked from popular apps and by malwares. Gibler et al. [13] presented AndroidLeaks, a static analysis framework for automatically finding potential leaks of sensitive information in Android applications on a massive scale. AndroidLeaks drastically reduces the number of applications and the number of traces that a security auditor has to verify manually. They found 57,299 potential privacy leaks in 7,414 Android applications, out of which they have manually verified that 2,342 applications leak private data. All these prior work tried to reduce the privacy risk with reference to some techniques but they did not show the exact level of risk. They did not show the risk of the entire PII map stored in mobile devices.

Different from the aforementioned work, we focus on privacy policies of most popular health and fitness apps in Goolge play store and iOS Apple store to see what set of PII these apps claim to collect in their privacy policies. We have a holistic list of over 600 PII. With the UT CID Identity Ecosystem, we are able to construct identity attribute maps that are generated from the mobile devices to answer questions that prior work could not answer. For example, “What is the total value of my PII stored in a single smart-phone?" or “What are the consequences if I loose my cellphone or it is stolen?”. Finally, we also ask “How do we better protect app user privacy?”.

3 METHODOLOGY

3.1 Experimental Setup

In this section, we introduce the source of our data set and mobile applications that we took under investigation.

3.1.1 ITAP Data. The Identity Threat Assessment and Prediction (ITAP) [28] is a research project at the Center for Identity that enhances fundamental understanding of identity processes, valuation, and vulnerabilities. The purpose of ITAP is to identify mechanisms and resources that are actually used to implement identity breach. ITAP cares about the exploited vulnerabilities, types of identity attributes exposed, and the impact of these events on the victims. Between years 2000 and 2018, about 6,000 incidents have been captured [1]. ITAP gathers details of media news stories (e.g., the PII attributes exposed, the location and date of the event, the age and annual income of the victims, and the perpetrators' methods) about identity theft with two methods. First, it monitored a number of Web sites that report on cases of identity theft. Second, it created a Google Alert to provide notifications when any new report of identity theft appears. By analyzing these cases, ITAP has generated a list of identity attributes with each of them being assigned identity-related vulnerabilities, values, risk of exposure, and other characteristics. To date, ITAP has generated a list including more than 600 identity attributes, which is the list of PII attributes we are referring to in this research.

3.1.2 Mobile Applications. The two biggest app stores are the Apple iOS App Store and Google Play store. We selected the top 100 most popular applications in the Apple App Store for iOS and the top 100 most popular apps in the Google Play Store for Android during March 2019. These 100 apps handle sensitive user data across different categories: Behaviors, Food and Drink, Interests, Social Networking, etc.

3.2 Privacy Policy

Privacy policies help users understand what portion of their sensitive data would be collected and used or shared by a specific mobile application. An app's privacy policy should be able to answer the following questions: What information does this application collect? How does this app use the information? and what information does this app share? A privacy policy should disclose all the information an app actively and passively collects, for example, information actively entered when registering for an account or passive HTTP logs and Internet usage. A privacy policy should disclose the purpose of collecting specific data. For example, the application might collect email addresses for promotion notification. Lastly, a privacy policy should disclose if the application shares information with any third parties like restaurants or ad trackers as part of the service.

We manually browsed the privacy policy of each of the 40 mobile applications to see what set of information they collect. Then we mapped the information against the ITAP identity attribute list.

4 FINDINGS

Our examination of popular smart healthcare apps for iPhone and Android phones—showed that around 35% (220 out of 627) of PII in the ITAP list are collected by these popular apps. We call this set of PII the examined set of PII. Each PII has their intrinsic properties, such as the attribute's type, its risk of exposure, and its monetary value. ITAP divided an attribute's type into four type of categories:

What You Are: a person's physical characteristics, such as biometrics like fingerprints and retina.

What You Have: documents and numbers assigned to a person by other organizations, such as a passport number.

What You Know: information known privately to a person, such as personal identification number (PIN).

What You Do: a person's behavior and action patterns, such as GPS location.

Figure 1 shows the portion that each type of PII has in the examined set of PII and Figure 2 shows the portion that each type of PII has in the ITAP list. The ITAP list is a comprehensive list of PII attributes collected from various cases in real life. We assume that the ITAP list is representative of all PII attributes. In the examined set, the type of "What You Know" and "What You Have" are the two largest proportion. That is, the mobile applications often want to collect information known privately by a person and information assigned by other organizations. The ITAP list has the same distribution.

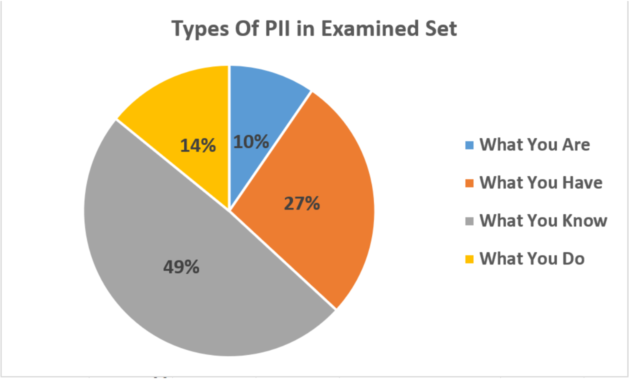

Fig. 1. The pie chart of the four types of PII for the examined set.

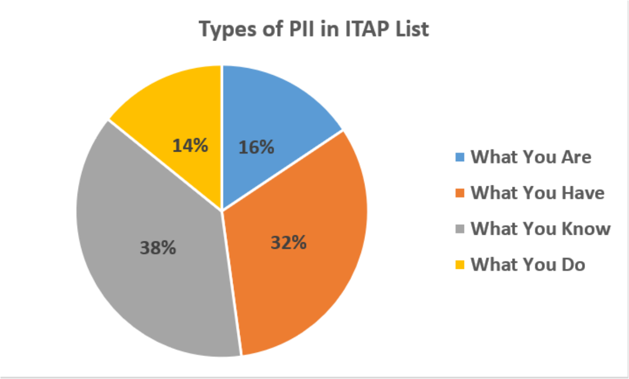

Fig. 2. The pie chart of the four types of PII for the ITAP set.

Compared with the ITAP list, mobile phones collect less PII attributes of the "What You Are" type. For example, PII attributes such as height and weight are usually collected by medical apps, but no medical app is in the top 20 most popular app. Therefore, the proportion of this type is relatively small. Even though the proportion is not significant, the type of "What You Are" is the

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

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

second highest percentage among comprised types of PII in 2018, while the type of "What You Have" is the highest [1].

In ITAP list, each identity attribute has been assigned a monetary value and risk of exposure that evaluated from ITAP resources. The average monetary value of the examined set of PII is around $5.4 million and the average risk of exposure is around 0.37%. In order to obtain a clear understanding of what the above numbers mean and put them in perspective, we include another set of PII for comparison: the PII in the U.S. Immigration and Citizenship Services (USCIS) Naturalization form. International migration has made significant contribution to overall American population growth during 2017 and 2018 [17]. Without a doubt, the U.S. Citizenship and Immigration Services has been dealing with tremendous amount of documentations, such as N-400 forms [27] which is used for Naturalization. The required identity attributes in the N-400 form are shown in Table 1. We call this set of PII the USCIS set of PII. The average monetary value of the USCIS set of PII is $2.65 million and the average risk of exposure is 3.2%. To produce an intuitive measure, we multiply the value and risk together to form the expected value. The expected value of the examined set is around 62k and the expected value of the USCIS set is around 85k. When people provide a set of PII to an organization, for example USCIS, the organization has the duty to keep the information safe. For immigration purposes, USCIS, which administers the country's naturalization and immigration system, collects a set of PII with the expected value of 85k and protects them with its protecting mechanisms. On the other hand, one's mobile phone may collect a set of PII with the expected value of 62k, which is around 73% of the USCIS set, but in a substantially less-protected environment. People are reluctant to provide PII to government agencies, but readily do so with their apps. To let mobile phones collect only what is necessary is what everyone should think about when using a mobile application.

In the examined set of PII, 41 of them, which is 6.5% of the ITAP list, are collected by all the popular apps. These PII attributes are often entered by users during the registration process or users just make them vincible on their social media. This active exposure of PII is a part of the users' *digital footprint*. Another part of the digital footprint is the data traces users leave behind passively, often referred to as *digital exhaust*. Figure 3 shows the proportion of PII automatically collected and manually entered in the examined set of PII. It shows that one third of the PII attributes are automatically collected by healthcare mobile applications, including GPS locations, interests, browsing history, etc.

Beyond mobile applications that are installed by users, some software is shipped with mobile devices. Such software is installed by the manufacturers and is called a device embedded system. The users generally cannot un-install such software, and often there is no accompanying privacy policy. As a result, we were not able to study the private policy of the embedded systems to see what information they collect, use, or share. In addition, mobile service providers, for example

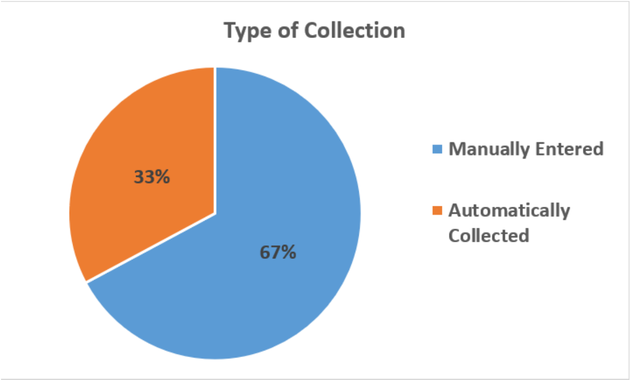

Fig. 3. The pie chart for type of collection for PII attributes.

AT&T, are also collecting PII attributes. Analyzing device embedded systems and service providers' collection and use of PII is an important future work.

In the examined set of PII, 53 out of 220, which is 8.5% of the 627 PII in the ITAP list, are originated from the physical world. Here we give the definition that helps us distinguish if the PII attribute is originated from physical world.

Definition 4.1. Originating from Physical World: A PII attribute is said to be originated from physical world if it can exist or be physically stored without any devices or digital storage.

Take the social security card for instance. When one first receives a social security card, it is a paper card. A social security card or number can exist even without any digital device or the Internet. Therefore, we say that a Social Security card is originated from the physical world. Compared with the past, when information was often recorded in the form of hard copies, today, with the rapid development of technology, virtually *all* information is stored in computers, mobile phones, and even the cloud. By analyzing PII that do or do not originate from the physical world, we see to distinguish between purely digital PII (which necessarily needs digital devices) and physical PII (which is not solely dependent on digital devices).

Another interesting finding is what we discovered from the set of PII collected only from medical apps. There is plenty of research into medical apps and previous work showed that medical apps collect certain amount of user sensitive data. They collect 11% of the ITAP list, which is considerable, whereas the set formed by the top 20 most popular apps collects 32% of the ITAP list. The top 20 apps spanned a range of categories such as social media, instant messenger, streaming service, etc. Hence, medical-related PII plays an important role in the ecosystem of one's identity. Different technologies are constantly being developed to improve users' health, and all types of sports and medical devices are constantly collecting user information. The connection between the device and the identity of the person and the analysis of the flow of personal information are major concerns of our work.

5 USE OF THE UT CID IDENTITY ECOSYSTEM

In this section, we cover our previous work, the UT CID Identity Ecosystem, a probabilistic model for relationship simulation, to demonstrate some analytics on the set of PII collected from popular apps.

5.1 The UT CID Identity Ecosystem

We first provide a high level introduction to our UT CID Identity Ecosystem [19]. The UT CID Identity Ecosystem developed at the Center for Identity at the University of Texas at Austin is a tool that models identity relationships, analyzes identity thefts and breaches, and answers several questions about identity management. It takes ITAP data as input and transforms them into identity attributes and relationships, and performs Bayesian network-based inference to calculate the posterior effects on each attribute. It presents identity attributes as nodes and various types of connections between nodes as edges. Each node has its own properties such as type of node, risk of exposure, and intrinsic monetary value.

Figure 4 shows a mini version example of the graphical model. Each node has a Boolean flag, a value and a prior risk of exposure. The Boolean flag "exposed" denotes whether the node is exposed or not. The intrinsic monetary value indicates the loss after the node is exposed. The prior denotes the probability of exposure of the node on its own. Note that the numbers in the graph are different in the real system.

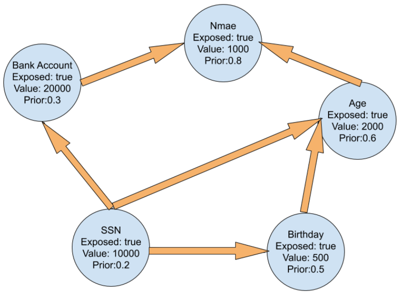

Fig. 4. Mini example of the UT CID Identity Ecosystem graphic model.

The Ecosystem Graphical User Interface (GUI) can color and size attribute nodes based on various properties. In Figure 5, nodes are colored based on their prior risk of exposure and are sized based on their value. Having this graphical model, we are able to answer some interesting question with the UT CID Identity Ecosystem.

5.2 Effect of the Exposure of PII Collected by Apps on Other PII

As explained above, in UT CID Identity Ecosystem, each PII attribute is represented by a graph node and various relationships between PII are shown as edges. Consequently, the UT CID Identity Ecosystem is capable of showing how the exposure of a PII attribute directly affects the exposure of another, wherein the latter is a child of the former in the graph.

The set of the first level children of the examined set of PII and the examined set itself together cover around 70% of the ITAP list. To clarify, for each attribute A_k, there is a set of out-degree edges denoted as $E_{out}(A_k)$ and the set of nodes directly connected with $E_{out}(A_k)$ is the set of first level children denoted as $CHILDREN_{first}(A_k)$. The total number of node of the examined set is about 35% of the ITAP list. In addition the examined set and $CHILDREN_{first}(A_k)$ (for A_k in the examined set) together cover 70% of the entire PII list. When a PII attribute is targeted by a perpetrator, its first level children are the most vulnerable. The edge between the two nodes by definition means that when this node is violated by a perpetrator, the next node that may be violated is one of its

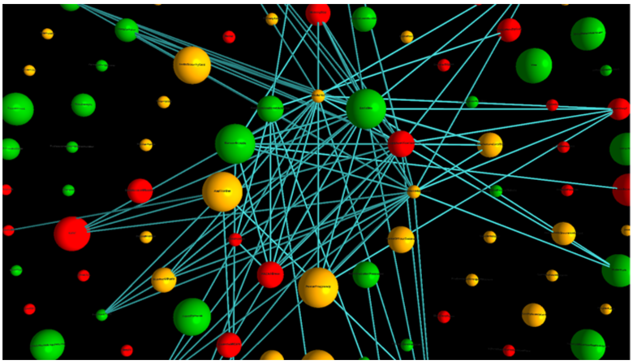

Fig. 5. A screen-shot of the UT CID Identity Ecosystem with nodes colored by properties.

first level children. Hence, the information users are sharing with mobile apps puts about 70% of their entire personally identifiable information at risk, directly or indirectly.

5.3 Analyzing the Risk of Exposure

We leverage one of the queries that the UT CID Identity Ecosystem supports, query 1, which infers probability of breach for other PII, based on the evidence that a given set of PII is compromised. To illustrate, imagine a user that accidentally accesses a suspicious website which asks for personal information for registration. Let us assume that the user reveals his/her social security number. After realizing it was a mistake, the user wants to know what risks this exposure imposes on his/her other identity attributes. Query 1 is precisely designed to answer such questions. By running query 1 in the UT CID Identity Ecosystem and providing social security number as initial evidence, the UT CID Identity Ecosystem shows the change in the probability of exposure of other PII after this incident. By looking at the results, the user is able to see what set of identity attributes are at high risk now. We are using query 1 on the set of PII collected by most popular apps. If the PII collected by these apps is compromised, e.g., the user has his/her phone stolen, how does it affect the risk of exposure of other PII for the user?

There are many healthcare applications being involved in this work so we group them together based on the number of identity attributes they collect. For each app, we look up the set of PII it collects from its privacy policy. Assuming that each set of PII is compromised and using it as evidence, we run query 1. Table 2 shows the number of PII collected, and the estimated potential cost when the set of PII this app collects is breached. Healthcare apps that provide food nutrition facts or fitness apps that are used for reminder purpose only are the most common type of app. They do not collect that much information from users. Most of them collect around 30 PII attribtues which includes user's basic information like password or birthday. the second large group of app goes for apps that track how many steps one has been walking or how far one has gone requires further information like gps location of the user or apps that diagnosed the user like mental health test would require more biometric data from users. This kind of app often collect from 60 to 90 PII attributes. The smallest group but collect most PII attributes is for apps that provides a platform for users to share ideas or their result of exercising. It can collect more than 100 PII attributes because it has its own social environment which can provide more information to industries. The most famous example is Facebook. Even thought it is not a healthcare app but it collects 132 PII attributes in ITAP list.

Table 2. Popular health and fitness apps in Android Google Play and Apple App Store and the cost of the exposure of the PII they collect.

Number of Apps	Number of PII	Exposure Cost ($)
38	~100	~1,000,000
60	60 ~90	~700,000
102	~30	~200,000

Table 3. 28 common PII attributes collected by popular apps.

Password	Username	Name	Address	Date of Birth
Phone Number	Account Number	Email Address	Zip Code	Gender
Login Credentials	Account Information	Age	Age Group	IP Address
Security Q&A	Time Stamp	Date	URL	Location
Mobile Device Data	Last Name	Last Login Date	Birth Year	Past Address
Initials	Postal Code	User Sign Up Date		

Apps that collect more than 100 PII attributes has higher potential monetary loss when its set of PII is breached. However, it is not always true that the higher the number of PII an app collects, the higher the cost of its possible breach. Only when the app collects important PII attributes the value of the whole set skyrockets. We observed that the set of PII required for registration purposes is one important set of PII. These apps collect a portion of the same PII attributes. We show all of them in Table 3. Many of these PII are often entered during registration.

5.4 Measuring the Accessibility and the Post Effect

To describe the importance of a specific PII attribute further, we leverage another two parameters we previously introduced [7], the Accessibility and the Post Effect, calculated by the UT CID Identity Ecosystem.

We call the first parameter *Accessibility.* In the calculation of a respective PII attribute's Accessibility, we analyzed the PII attribute's ancestors (in the UT CID Identity Ecosystem graph) to assess the probability and likelihood of discovering this PII node (attribute) from other nodes. These "discovery" probabilities on edges in the UT CID Identity Ecosystem graph are calculated using UT CID ITAP data representing how criminals discovered PII attributes using a respective PII attribute. Low values of Accessibility indicate that it is more difficult to discover to this attribute from others. A PII attribute with low Accessibility is harder to breach or discover (discoverability). Figure 6 shows the Accessibility measures for those PII attributes collected by the mobile applications under investigation. The attribute "Name" has the highest Accessibility, which makes no surprise since a person is often known by his or her name first. The PII attribute with the second highest accessibility is the "Address" which is often found in public records. Popular apps also collect PII attributes with very low accessibility like "IP_Address" and "Location".

We call the second parameter *Post Effect.* For a target PII attribute, we analyze the PII attribute's descendants in the UT CID Identity Ecosystem graph. If a PII attribute is breached, the Post Effect measure gauges how much the respective PII attribute would influence other attributes. The low value of Post Effect of an attribute indicates that the damage or loss one would encounter is smaller after this PII attribute is accessed by fraudsters. Figure 7 shows the Post Effect measures for those PII attributes collected by the mobile apps under investigation. The attribute with the highest Post Effect is "Password" which gives access to further sensitive information like social media accounts

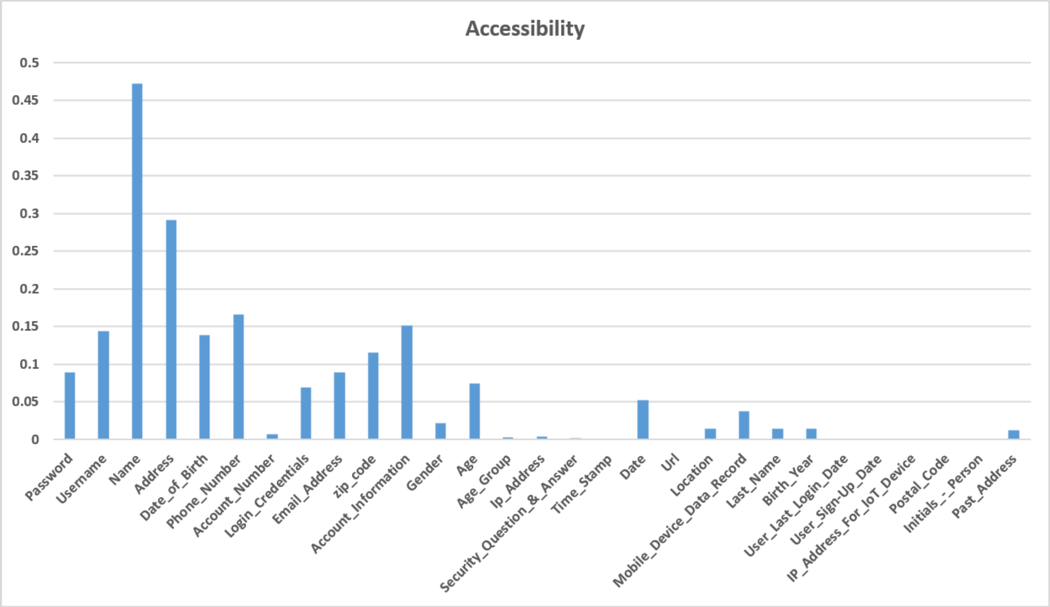

Fig. 6. The bar chart for Accessibility.

and bank account details. The PII attribute with the second highest Post Effect is the "Location". On the other hand, the attribute "Name" has lower Post Effect than others because names have already been exposed and attributes like "Phone_Number" and "Age" are similar to "Name" in their value of Post Effect as they are in public records.

Comparing Accessibility and Post Effect together could give us some clear insights. Figure 8 shows the scatter diagram with the Post Effect on the vertical axis and the Accessibility on the horizontal axis. According to the diagram, PII attributes that have higher Accessibility often have lower Post Effect and vice versa. The evaluation of these two parameters is a step toward understanding how much PII attributes collected by mobile applications are important to privacy risks associated with the use of our IoT devices.

6 CONCLUSION

In this paper, we sought to understand the set of Personal Identifiable Information (PII) collected, used and shared by health and fitness mobile applications. Whether the applications are installed by the phone manufacturer or downloaded by the user, each app has a privacy policy explaining how that app collects, uses and protects PII. This research analyzed those privacy policies to begin to uncover just how much of our PII is on our phones and potentially shared/traded on the internet. Our experiment evaluated the privacy policies of 200 popular mobile apps. Our experiment compared the PII collected from these mobile apps to a reference list of over 600 PII attributes collected in the Identity Theft Assessment and Prediction (ITAP) project at The University of Texas. The ITAP project investigates theft and fraud user stories to assess how PII is monetized and the risk (likelihood) of respective PII attributes to be stolen and/or fraudulently used. From these mobile apps, our results indicate that 35% of the over 600 reference PII attributes were being collected.

If PII attributes are classified as "What you KNOW", "What you HAVE", "What you ARE", and "What you DO", we found that almost half the PII attributes (49%) collected by the mobile apps fall

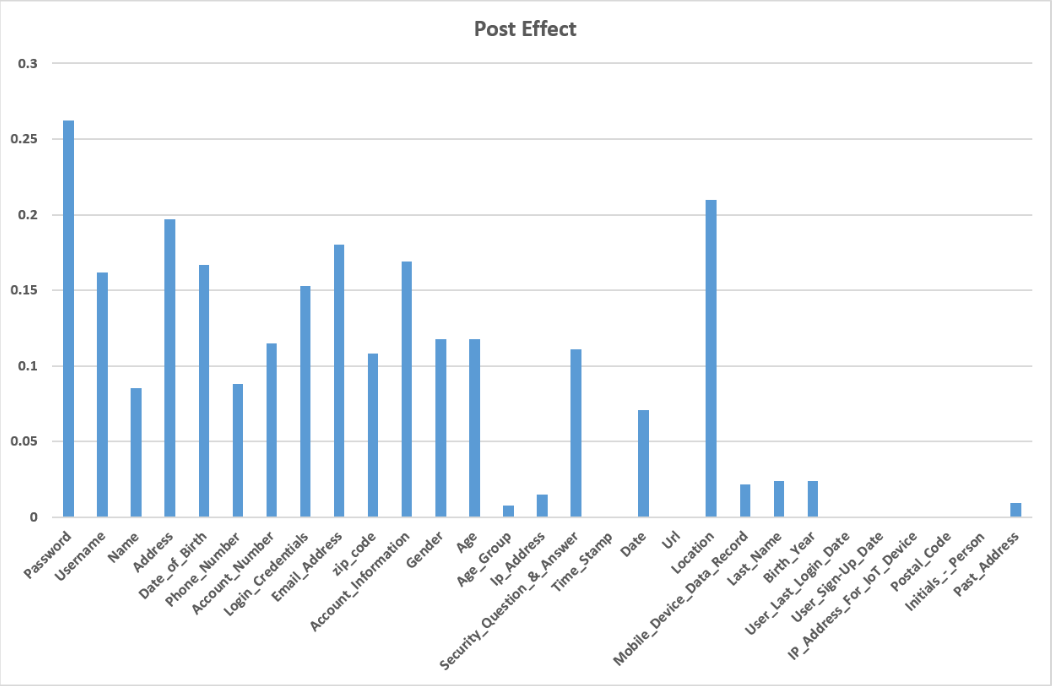

Fig. 7. The bar chart for Post Effect.

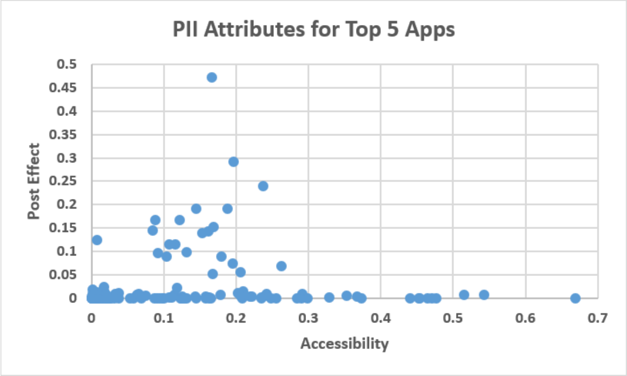

Fig. 8. The scatter diagram with post effect as vertical axis and accessibility as horizontal axis.

into the category of "What You KNOW" and 27% of the type "What you HAVE." The "What you DO" PII attributes only accounted for 10% of personal data collected by the mobile apps. This is noteworthy since it is an indicator of the behavioral surveillance conducted by mobile apps and IoT devices. Further investigation is needed to explore the degree of user awareness and consent as well as a longitudinal look at the increase over time as convenience offered via the collection of these behavioral biometrics trumps the security risks.

We also made comparisons between the set of PII attributes collected by mobile apps and the set of PII attributes used for immigration purposes at USCIS. The average expected value of breach cost for the studied mobile apps' PII set is around 62k while it is about 85k for the USCIS set. Yet, in comparison with USCIS, people are reluctant to provide PII to government agencies, but readily do so with their apps.

We also took advantage of two parameters, accessibility (discoverability) and post effect (potential impact on other PII attributes), calculated by the UT CID Identity Ecosystem. The higher a PII attribute's accessibility value, the more difficult it is to access or discover this respective PII attribute. If a PII attribute has a higher post effect measure, the PII attribute will have a higher impact on and likelihood to expose other PII atttributes. The attribute "Name" has the highest accessibility measure, and PII attribute with the second highest accessibility measure is the "Address" which is often in public records. The attribute with the highest post effect is "Password" and the PII attribute with the second highest post effect is the "Location". PII attributes from the five most popular apps have a wide range of accessibility representing a wide range of discoverability and consequently likelihood for privacy risks. While a smaller percentage (13%) of the PII collected have a high post effect, these key PII attributes will have high and direct impact on other PII if exposed.

This work was the first to study privacy policies of health and fitness mobile apps in terms of the PII collected, used and shared then further study those PII attributes in the context of a personal data reference model built by the UT CID Identity Ecosystem and ITAP projects. This research found that popular mobile apps put 35% of a user's personal data directly at risk and, due to dependencies between PII attributes, these mobile apps indirectly put an additional 35% or a total of 70% of a user's over 600 reference PII attributes at risk.

To answer the question "Is Your Phone You?", we conclude the answer is "yes." If these mobile apps could represent 70% of the reference PII attributes (over 420 PII attributes) studied and it takes far fewer PII attributes to claim an identity then, a user's phone has all the information required to assert a user's identity in the Internet of Things. This result is a warning sign for user privacy in the Internet of Things for one of its most common "things", a smart-phone.

REFERENCES

[1] Itap report 2018. Technical report, Center for Identity, University of Texas at Austin, 2018.

[2] L. Ackerman. Mobile health and fitness applications and information privacy. Technical report, Privacy Rights Clearing House, San Diego, CA, 2013.

[3] Y. Agarwal and M. Hall. Protectmyprivacy: Detecting and mitigating privacy leaks on ios devices using crowdsourcing. pages 97–110, 06 2013.

[4] E. Anthi and G. Theodorakopoulos. Sensitive data in smartphone applications: Where does it go? can it be intercepted? In X. Lin, A. Ghorbani, K. Ren, S. Zhu, and A. Zhang, editors, *Security and Privacy in Communication Networks*, pages 301–319, Cham, 2018. Springer International Publishing.

[5] S. Aslam. Facebook by the numbers: Stats, demographics & fun facts, jan 2019.

[6] R. Balebako, J. Jung, W. Lu, L. F. Cranor, and C. Nguyen. "little brothers watching you": Raising awareness of data leaks on smartphones. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, SOUPS '13, pages 12:1–12:11, New York, NY, USA, 2013. ACM.

[7] K. C. Chang, R. N. Zaeem, and K. S. Barber. Enhancing and evaluating identity privacy and authentication strength by utilizing the identity ecosystem. In *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, pages 114–120. ACM, 2018.

[8] T. Dehling, A. Sunyaev, P. L. Taylor, and K. D. Mandl. Availability and quality of mobile health app privacy policies. *Journal of the American Medical Informatics Association*, 22(e1):e28–e33, 08 2014.

[9] C. DOrazio and K. R. Choo. A generic process to identify vulnerabilities and design weaknesses in ios healthcare apps. In *2015 48th Hawaii International Conference on System Sciences*, pages 5175–5184, Jan 2015.

[10] M. Egele, C. Kruegel, E. Kirda, and G. Vigna. PiOS : Detecting privacy leaks in iOS applications. In *NDSS 2011, 18th Annual Network and Distributed System Security Symposium, 6-9 February 2011, San Diego, CA, USA*, San Diego, UNITED STATES, 02 2011.

[11] W. Enck, P. Gilbert, S. Han, V. Tendulkar, B.-G. Chun, L. P. Cox, J. Jung, P. McDaniel, and A. N. Sheth. Taintdroid: An information-flow tracking system for realtime privacy monitoring on smartphones. *ACM Trans. Comput. Syst.*, 32(2):5:1–5:29, June 2014.

[12] V. Fedorychak. Interesting mobile app market statistics for 2019, jan 2019.

[13] C. Gibler, J. Crussell, J. Erickson, and H. Chen. Androidleaks: Automatically detecting potential privacy leaks in android applications on a large scale. pages 291–307, 06 2012.

[14] S. Han, J. Jung, and D. Wetherall. A study of third-party tracking by mobile apps in the wild. Technical report, University of Washington, 2012.

[15] K. D. Harris. Privacy on the go. Technical report, California Department of Justice, 2013.

[16] P. Hornyack, S. Han, J. Jung, S. Schechter, and D. Wetherall. These aren't the droids you're looking for: Retrofitting android to protect data from imperious applications. In *Proceedings of the 18th ACM Conference on Computer and Communications Security*, CCS '11, pages 639–652, New York, NY, USA, 2011. ACM.

[17] J. Kolko. How much slower would the u.s. grow without immigration? in many places, a lot. Technical report, The New York Times, 2019.

[18] C. Liu and K. P. Arnett. An examination of privacy policies in fortune 500 web sites. *American Journal of Business*, 17(1):13–22, 2002.

[19] R. Nokhbeh Zaeem, S. Budalakoti, K. S. Barber, M. Rasheed, and C. Bajaj. Predicting and explaining identity risk, exposure and cost using the ecosystem of identity attributes. In *2016 IEEE International Carnahan Conference on Security Technology (ICCST)*, pages 1–8. IEEE, 2016.

[20] A. Papageorgiou, M. Strigkos, E. Politou, E. Alepis, A. Solanas, and C. Patsakis. Security and privacy analysis of mobile health applications: The alarming state of practice. *IEEE Access*, 6:9390–9403, 2018.

[21] A. Raoa, A. M. Kakhkib, A. Razaghpanahe, A. Tangc, S. Wangd, J. Sherryc, P. Gille, A. Krishnamurthyd, A. Legouta, A. Misloveb, and D. Choffnesb. Using the middle to meddle with mobile. Technical report, Northeastern University, 2013.

[22] M. Rowan and J. Dehlinger. A privacy policy comparison of health and fitness related mobile applications. *Procedia Computer Science*, 37:348 – 355, 2014. The 5th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN-2014)/ The 4th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH 2014)/ Affiliated Workshops.

[23] E. Smith. iphone applications & privacy issues: An analysis of application transmission of iphone unique device identifiers (udids). Technical report, 2010.

[24] J. Stern. iphone privacy is broken…and apps are to blame, may 2019.

[25] SWNS. Americans check their phones 80 times a day: study, nov 2017.

[26] S. Thurm and Y. Kane. Your apps are watching you. *Wall Street Journal*, 2010.

[27] uscis. U.s. citizenship and immigration services.

[28] J. Zaiss, R. Nokhbeh Zaeem, and K. S. Barber. Identity threat assessment and prediction. *Journal of Consumer Affairs*, 53(1):58–70, 2019.

[29] S. Zhao, X. Luo, B. Bai, X. Ma, W. Zou, X. Qiu, and M. H. Au. I know where you all are! exploiting mobile social apps for large-scale location privacy probing. In J. K. Liu and R. Steinfeld, editors, *Information Security and Privacy*, pages 3–19, Cham, 2016. Springer International Publishing.

WWW.IDENTITY.UTEXAS.EDU