The University of Texas at Austin
Center for Identity

A Framework for Estimating Privacy Risk Scores of Mobile Apps

Kai Chih-Chang
Razieh Nokhbeh Zaeem
K. Suzanne Barber

UTCID Report #20-11

June 2020

A Framework for Estimating Privacy Risk Scores of Mobile Apps

Kai Chih Chang[1][0000−0002−9307−2358], Razieh Nokhbeh Zaeem[1][0000−0002−0415−5814], and K. Suzanne Barber[1][0000−0003−2906−6583]

The University of Texas at Austin, Austin TX 78712, USA
{kaichih,razieh,sbarber}@identity.utexas.edu

Abstract. With the rapidly growing popularity of smart mobile devices, the number of mobile applications available has surged in the past few years. Such mobile applications collect a treasure trove of Personally Identifiable Information (PII) attributes (such as age, gender, location, and fingerprints). Mobile applications, however, are many and often not well understood, especially for their privacy-related activities and functions. To fill this critical gap, we recommend providing an automated yet effective assessment of the privacy risk score of each application. The design goal is that the higher the score, the higher the potential privacy risk of this mobile application. Specifically, we consider excessive data access permissions and risky privacy policies. We first calculate the privacy risk of over 600 PII attributes through a longitudinal study of over 20 years of identity theft and fraud news reporting. Then, we map the access rights and privacy policies of each smart application to our dataset of PII to analyze what PII the application collects, and then calculate the privacy risk score of each smart application. Finally, we report our extensive experiments of 100 open source applications collected from Google Play to evaluate our method. The experimental results clearly prove the effectiveness of our method.

Keywords: Mobile Applications · Privacy · Privacy Policy · Permissions · Natural Language Processing

1 Introduction

In recent years, portable smart devices have rapidly spread, bringing a large number of mobile applications to various users. For example, as of May 2020, Google Play has more than 3 million Apps, which is three times the number in 2013, and these numbers are still growing rapidly. Due to the prosperous development of the smart application industry, the functions of smart devices have been extensively expanded and innovated to meet the needs of diverse users. However, the types of mobile applications are ever-changing, and their contents and architecture are often difficult to understand. Questions about their activities and functions related to privacy and security are endless. In fact, in order to improve the user experience, more and more advanced mobile applications

are inclined to gather user data to provide personalized service. These services usually involve access to sensitive personal information such as location.

However, such intelligent mobile Apps may result in potential security and privacy risks for users. So much Personally Identifiable Information (PII) is hidden in a smartphone, such as What We Are (e.g., fingerprints), What We Have (e.g., credit card information), What We Know (e.g., email password), and What We Do (e.g., location history)[29]. We call these PII attributes identity assets. In addition, emerging technologies of IoT (Internet of Things) bring new forms of user interfaces, such as wearable devices, which also pose greater challenges to user privacy. Therefore, it is important to study what identity assets are collected by these mobile applications.

A privacy policy is one of the most common methods of providing user notifications and choices. The purpose of a privacy policy is to inform users how the application collects, stores and discloses users' identity assets. Although some service providers have improved the intelligibility and readability of their privacy policies, not everyone reads them. As of 2019, only 24% of people read the privacy policy [15].

Another potential privacy risk for mobile applications is basically caused by excessive data access permissions of mobile applications. As mentioned earlier, the current mobile applications provide a variety of innovative services, and these services involve various data access permissions. Sometimes these permissions are necessary, sometimes not. Therefore, in this paper we propose to leverage the requested permissions and privacy policies for detecting the potential privacy risk of each mobile App.

To create a comprehensive list of PII, we utilize our longitudinal study of 6,000 identity theft and fraud news stories reported over the past 20 years. This database–named Identity Threat Assessment and Prediction or ITAP [31, 30]– is a structured model of PII, manually extracted by a team of modelers from identity theft and fraud reports in the online news media. We take advantage of ITAP to evaluate the risk score of each identity asset in order to estimate the privacy risk score of the set of identity assets that a mobile App collects.

This paper makes the following contributions:

1. We map an independently built, comprehensive list of identity assets to privacy polices and data access permissions in order to evaluate the privacy risk score of mobile apps.
2. We use Natural Language Processing (NLP) methods to automatically parse privacy policies to find the identity assets mentioned in them.
3. Having access to UT CID probabilistic models and Bayesian inference tool Ecosystem [21], we take advantage of Bayesian inference to help calculate privacy risk score of mobile apps.
4. We demonstrate how our approaches can work on 100 popular open-source Android mobile Apps in Google Play and compare our results to other researchers' work.

2 Data and Methodology

In this section, we briefly introduce the dataset that we are using and also the details of our privacy risk measurements.

2.1 UT CID ITAP Dataset

The Identity Threat Assessment and Prediction (ITAP) [31, 30] is a research project at the Center for Identity at the University of Texas at Austin 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 2019, about 6,000 incidents have been captured [3]. ITAP gathers details of media news stories (e.g., the identity assets 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 depending on their properties, such as, whether or not an attribute is unique to a person, whether or not an attribute is widely used, how accurately it can be verified, etc. To date, ITAP has generated a list including over 600 identity assets, which is the list of identity assets we are referring to in this research.

Each identity asset in the UT CID ITAP dataset has a group of properties, including, but not limited to the following properties:

Risk: indicates the probability of this identity asset being misused in identity theft and fraud incidents.

Value: indicates the monetary value of this identity asset when misused in identity theft and fraud incidents.

2.2 Identity Assets Collection from Apps

Privacy risks are essentially caused by the data collections of Apps. Thus, an intuitive approach for measuring the privacy risks of Apps is to directly check each of the identity asset they collect/request. In this work, we divide data collection into two parts: (1) the privacy policy of each apps and (2) the Android manifest XML file of each apps.

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. By reading the privacy policy of an app, we should know what information

this application collects, how this app uses the information, and what information this app shares. A privacy policy discloses 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.

The *bag-of-words* (BoW) model is a simplifying representation used in natural language processing and information retrieval. We construct a BoW model and take the privacy policy as input to generate a list of words and map it to the ITAP dataset to see what identity assets this privacy policy collects. In our model, we manually map each *word* to different identity assets so that after feeding our model with the privacy policy, we can generate a set of identity assets that this app collects. Table 1 shows some example of BoW mapping. We define the set of identity assets of app S that includes N identity assets as

$$Set_{BoW}(S) = \{x_i\}_{i=1:N} \tag{1}$$

where x_i denotes the identity asset in UT CID ITAP dataset.

Table 1. Examples of privacy policies mapping to ITAP dataset.

Words	Correlated Identity Assets
Email	Email_Address
Name	User_Name
Phone	Phone_Number
Location	GPS_Location

XML File To access the personal data in users' Android mobile devices, the permission system will convey users to grant corresponding data access permissions for each mobile app. Actually, these data access permissions may enter some sensitive resources in mobile users' personal data, such as their locations or contact lists. Table 2 shows some example of permissions. We can see that these listed permissions contain potential security risks. For example, an App, which requests READ_CALENDAR permission, may access users' personal calendar which could make users like businesspersons feel uncomfortable due to leaking their schedules. In this work, we construct a program in which we manually map each Android permission to identity assets in UT CID ITAP dataset. This program takes Android manifest file as input and generate a set of identity assets that this app collects. Table 3 shows some mapping example of permissions. We define the set of identity assets of app S that includes N identity assets as

$$Set_{XML}(S) = \{x_i\}_{i=1:N} \tag{2}$$

where x_i denotes the identity asset in UT CID ITAP dataset.

Therefore, we can define a dataset of identity assets for app S as

$$ID_S = Set_{BoW}(S) \cup Set_{XML}(S) \tag{3}$$

Table 2. Examples of Android permissions.

Type	Permission Name	Description
String	ACCESS_BACKGROUND_LOCATION	Allows an app to access location in the background.
String	NFC_TRANSACTION_EVENT	Allows applications to receive NFC transaction events.
String	READ_CALENDAR	Allows an application to read the user's calendar data.
String	READ_CALL_LOG	Allows an application to read the user's call log.

Table 3. Examples of Android permissions mapping to ITAP dataset.

Permission Name	Correlated Identity Assets
ACCESS_BACKGROUND_LOCATION	GPS_Location
NFC_TRANSACTION_EVENT	Transaction_Records
READ_CALENDAR	Calendar_Information
READ_CALL_LOG	Call_History

2.3 Estimating Risk Scores For Identity Assets

Generally speaking, the risk score should reflect the security level of an identity asset. The higher the score is, the more dangerous when the identity asset is exposed. Dangerous here means the danger of monetary loss one could have encountered when the identity asset of this person is exposed. Recall that ITAP associates monetary values to identity assets.

We have two approaches to calculate the risk score of identity assets. Among those properties, we first choose *risk* and *value* for measuring the risk score of each identity asset.

Basic Measurement Given N identity assets in UT CID ITAP dataset, each identity asset A_i is labeled with a monetary value $V(A_i)$ and a prior probability $P(A_i)$ of it getting exposed on its own. We define the expected loss of an identity asset A_i as

$$Exp(A_i) = P(A_i) \cdot V(A_i) \tag{4}$$

such that $1 \leq i \leq N$.

Dynamic Measurement We have another way for calculating expected loss. Instead of using only intrinsic values of identity assets in UT CID ITAP dataset, we leverage two more parameters which we introduced in previous work [6] to refine risk and value of identity assets.

We first provide a high level introduction to our UT CID Identity Ecosystem [21, 6, 24, 7, 8, 18]. 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 UT CID ITAP dataset as input and transforms them into identity assets and relationships, and performs Bayesian network-based inference to calculate the posterior effects on each attribute. We represent UT CID Identity Ecosystem as a graph $G(V, E)$ consisting of N identity assets $A_1, ..., A_N$ and a set of directed edges as a tuple $e_{ij} =< i, j >$ where A_i is the originating node and A_j is the target node such that $1 \leq i, j \leq N$. Each edge e_{ij} represents a possible path by which A_j can be breached given that A_i is breached.

The first parameter we reuse from our previous work is called *Accessibility*. In the calculation of a respective identity asset's accessibility, we analyzed its ancestors (in the UT CID Identity Ecosystem graph) to assess the probability and likelihood of discovering this node from other nodes. These "discovery" probabilities on edges in the UT CID Identity Ecosystem graph are calculated using UT CID ITAP dataset representing how criminals discovered identity assets using a respective identity asset. Low values of accessibility indicate that it is more difficult to discover to this attribute from others. An identity asset with low accessibility is harder to breach or discover (discoverability). Since accessibility is the change in risk of exposure, we can calculate new risk of an identity asset A_i as

$$P'(A_i) = P(A_i) + AC(A_i) \tag{5}$$

where $AC(A_i)$ denotes the accessibility of A_i.

The second parameter we obtain from our previous work is called *Post Effect*. For a target identity asset, we analyze its descendants in the UT CID Identity Ecosystem graph. If an identity asset is breached, the post effect measure gages how much the respective identity asset would influence others. The low value of post effect of an attribute indicates that the damage or loss one would encounter is smaller after this identity asset is accessed by fraudsters. Since post effect is also the monetary value, we can calculate new value of an identity assets A_i as

$$V'(A_i) = V(A_i) + PE(A_i) \tag{6}$$

where $PE(A_i)$ denotes the post effect of A_i.

Hence, for dynamic measurement, we define expected loss of identity asset A_i as

$$Exp(A_i) = P'(A_i) \cdot V'(A_i) \tag{7}$$

Since the range of the expected loss in UT CID ITAP dataset is from 0 to 10^7, which is quite wide, in order to rank each identity asset based on expected loss, we apply natural logarithm on each identity asset's expected loss which can be shown as $\ln(Exp(A_i))$. As we mentioned at the beginning of this section, the higher the score is, the more dangerous when the identity asset is exposed. To achieve this goal, we find the maximum value of expected loss after applying natural logarithm and use it to calculate the risk score of each identity asset. Thus, we define the risk score of an identity asset A_i as

$$score_{risk}(A_i) = \frac{ln(Exp(A_i))}{Max} \tag{8}$$

where Max denotes the maximum value of expected loss after applying natural logarithm. Hence, the risk score becomes a value that is normalized between 0 and 1.

2.4 Ranking For Mobile Apps

Then, we can compute risk scores of mobile apps with risk scores of identity assets. Given an app S that collects N identity assets, by our data collection approach, we can derive an identity asset dataset $ID_S = \{x_i\}_{i=1:N}$. For all of the members of app S, we can estimate the total risk score of the collected dataset:

$$Privacy_s = \frac{1}{Total} \sum_{i=1}^{N} score_{risk}(A_i) \tag{9}$$

where $Total$ denotes the sum of risk score of the entire UT CID ITAP dataset. Thus, the privacy risk score becomes a value that is also normalized between 0 and 1.

Therefore, we can also calculate the privacy risk score of one's mobile devices by adding up privacy risk scores of apps that one's device have installed.

3 Experimental Results

In this section, we empirically evaluate our app privacy ranking approaches with real-world Android apps.

3.1 Experimental Apps

In order to perform data collection analysis on manifest XML files, we target Android apps that are open-source. We found 100 Android apps that have privacy policies on Google Play and the source code of each of them is available on GitHub. Most of them are still actively maintained. Fig. 1 illustrates some statistics of the application dataset. It shows the number of Apps and the average number of requested permissions by each App in different categories. In this figure, we can observe that Apps in categories "Communication", "Business" and "Travel & Local" request more permissions and that we have more Apps in categories "Tools" and "Productivity" in this dataset.

3.2 Evaluation of App Privacy Risk Scores

Here, we evaluate the effectiveness of estimating App risk scores and compare our methodologies with previous work.

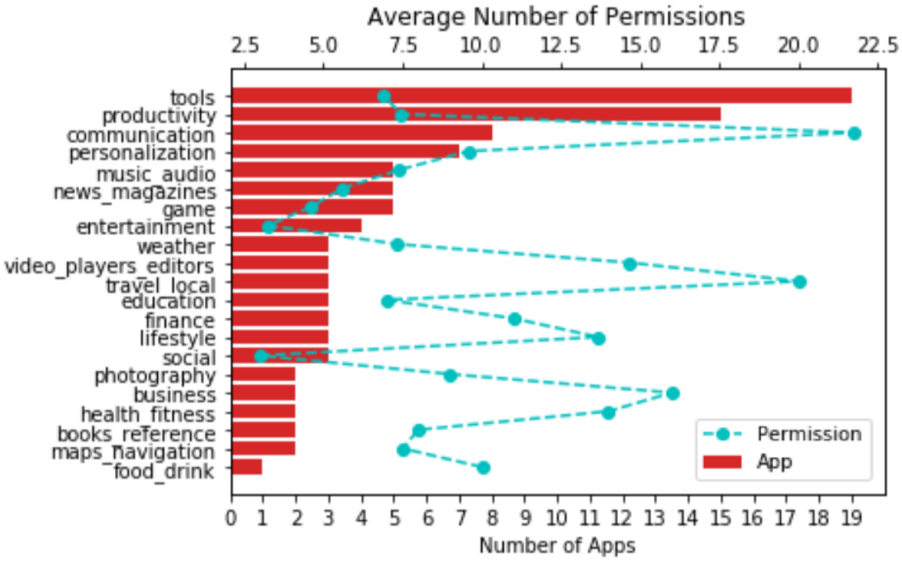

Fig. 1. The number of Apps and the average number of requested permissions by each App in different categories.

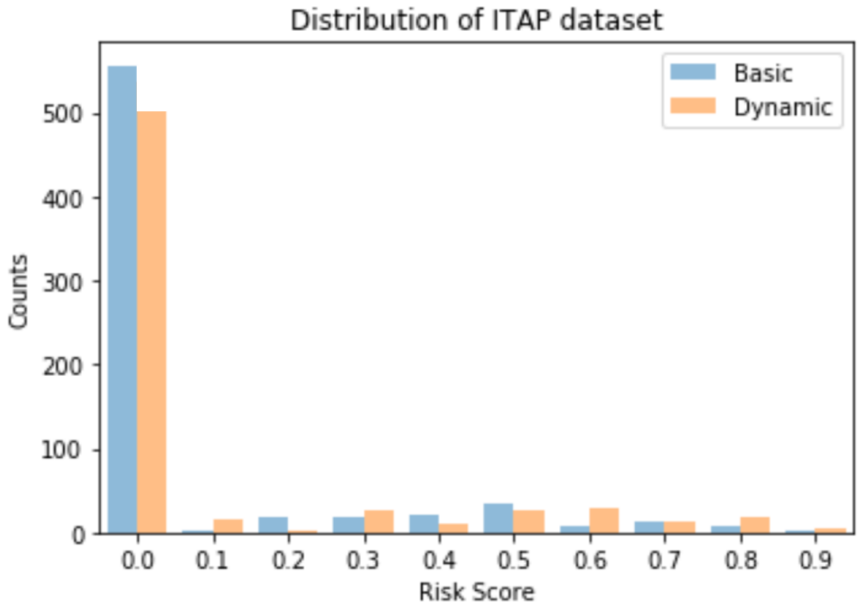

Fig. 2. The value of each rank and the number of identity assets with that rank.

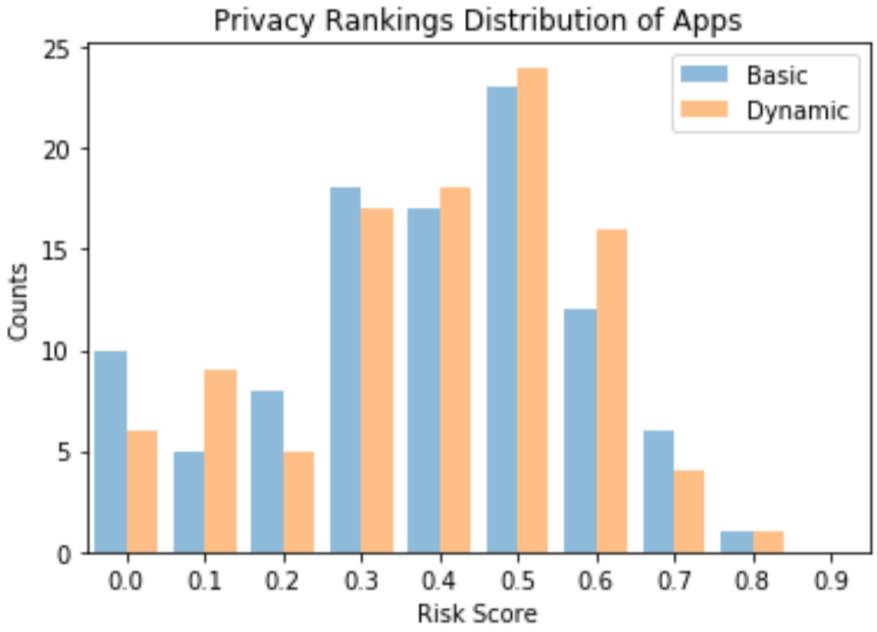

Fig. 3. The ranking distribution of basic and dynamic measurements on Android Apps.

General Results Fig. 2 shows the histogram of how many identity assets have a given rank value, according to both basic and dynamic methods of calculation. There are many identity assets that have the monetary value 0 reported from ITAP, because the monetary loss of the identity asset's exposure was not reported in the ITAP news stories. As a result, the number of identity assets in the lowest rank is relatively higher than the rest of ranks. As we mentioned in the methodology section, we apply the dynamic method in order to refine value and risk of identity assets. The dynamic measurement has reduced around 10% (50 identity assets) of the number of identity assets in the lowest rank and those 10% of identity assets have spread into different ranks due to their accessibility and post effect.

Fig. 3 shows the score distribution of the experimental Apps. In this figure, we observe that it has lots of numbers concentrated in the middle of the range, with the remaining numbers trailing off on both sides which is close to a normal distribution. The average risk score of the experimental dataset is 0.4469 or 44.69%. The identity asset that has highest risk score (which means it is most dangerous in the ITAP dataset) according to both approaches is "Social Security Number".

Like what we did in Fig. 1, we also analyze risk score with different App categories. Fig. 4 shows the average score of different categories of basic and dynamic measurements. From Fig. 1 we know that category of "Communication", "Business" and "Travel & Local" request more permissions and these categories

also have the highest average scores in Fig. 4. Also, category of "Weather" and "Food & Drink" do not request many permissions but are still in the higher tier of average score. On the other hand, in Fig. 5, it shows the correlation between risk score of apps and number of permissions they request. Even though not dramatically, according to the regression lines, when the number of permissions increases, the value of privacy risk score slightly increase as well.

Last but not least, we map identity assets in ITAP dataset to both privacy policy and XML file. Overall, the entire experimental dataset collects 70% of identity assets in the ITAP dataset while privacy policies collect 67% of identity assets and XML files only have 10% of identity assets which makes sense since we parse the entire privacy policy to map identity assets and meanwhile the maximum number of permissions that an app would request is only 32.

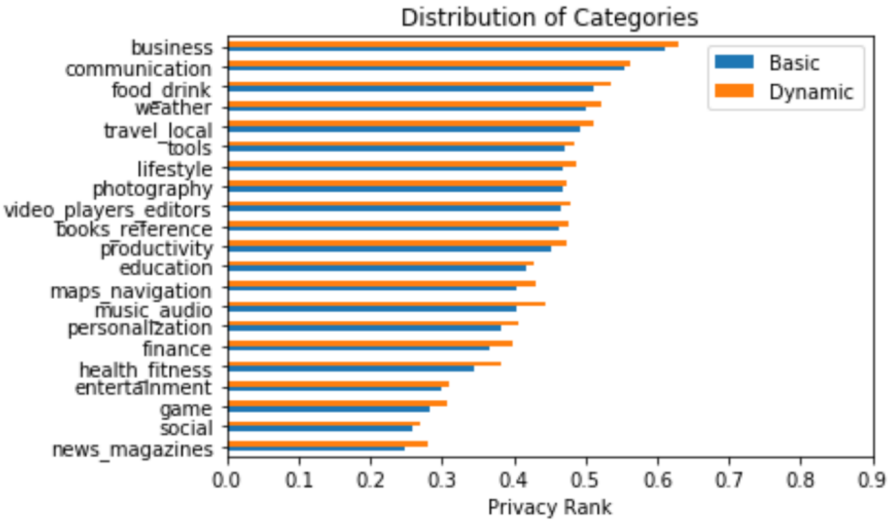

Fig. 4. The average score of different categories of basic and dynamic methods.

Evaluation of Ranking App Risk We adopt two baselines to evaluate the effectiveness of our approaches in terms of ranking App risks. The first work was introduced in 2019 by O'Loughlin et. al [22]. They evaluated the presence and quality of a privacy policy of apps with questions that aim to assess comprehensiveness of an app's documentation in describing data collection and storage practices and policies. By answering their questions in their work, they divided the score of the privacy policy into three ranks: "Acceptable", "Questionable", and "Unacceptable". In this section, we denote this approach as "OLoughlin".

The other tool that we use in this comparison as baseline is the ImmuniWeb®Mobile App Scanner [2] (short for ImmuniWeb). It is a tool that develops

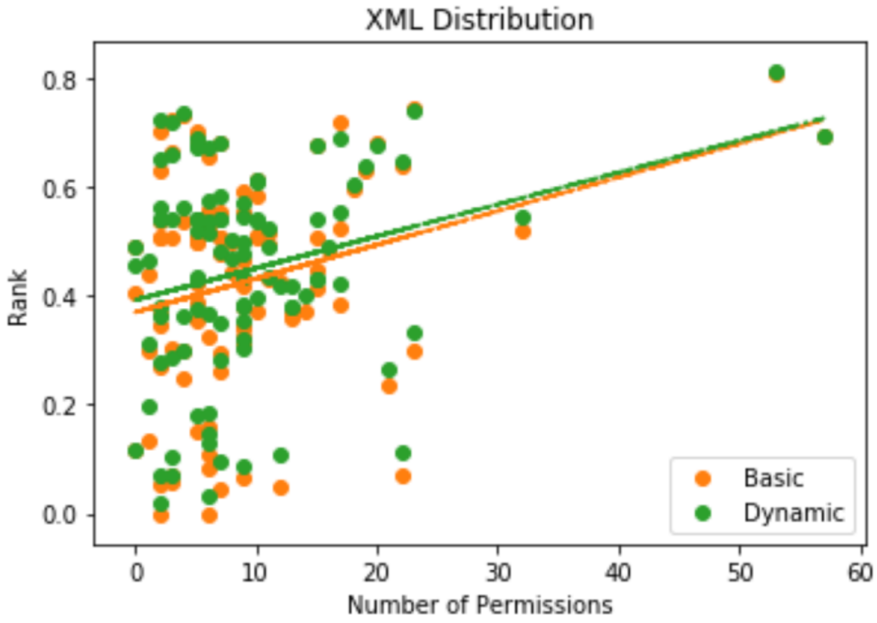

Fig. 5. The scatter diagram of number of permissions and risk score.

Machine Learning and Artificial Intelligence technologies for Application Security Testing and Attack Surface Management. Their automated tests reveal several security risk flaws and weaknesses that may impact the application. We pick tests that are related to privacy and data access like *Exposure of potentially sensitive data*. The level of each risk that has been detected can be divided into four ranks: "High", "Medium", "Low", and "Warning". "High" denotes the red light which indicates that this App has higher risk with respect to the according weakness or flaw.

We pick the most popular apps in our experimental dataset to compare our dynamic approach to different measurements. Each of the popular apps has over 5 million downloads in Google Play [1]. Table 4 shows the value of each popular App returned by each approach. The table is sorted by the value of our dynamic approach. We can see that almost every app in the first half of the table are being labeled as "Low" in ImmuniWeb. First 5 apps also have higher risk scores than others in the table. Therefore, we can see that our measurement is promising. The interesting thing is that in OLoughlin, as long as the privacy policy of this app does not mention whether its server encrypts users' information or not, this app is labeled as "Unacceptable". *Duckduckgo* and *OpenVPN*, which are located in the middle of the table, are the only two apps that are labeled as "Acceptable" in OLoughlin.

Table 4. The popular open-source Apps.

App	Dynamics(%)	ImmuniWeb	OLoughlin
Wiki	43.63	Low	Unacceptable
Firefox Focus	47.99	Low	Questionable
Kodi	48.79	Low	Unacceptable
QsmAnd	54.51	Low	Questionable
Duckduckgo	67.39	Medium	Acceptable
OpenVPN	68.92	Medium	Acceptable
Signal Private Messenger	69.32	Medium	Questionable
Ted	71.82	Low	Questionable
Blockchain Wallet	73.67	Medium	Questionable
Telegram	73.99	Medium	Questionable

4 Related Work

Generally speaking, research on mobile privacy risk can be divided into three categories: mobile App's permission analysis, mobile App's privacy policy analysis, and mobile security and privacy framework.

For the first category, mobile App's permission analysis, several works have been published. More and more mobile applications are providing novel services by requesting bunch of access permissions of user's sensitive information. To understand this, for example, Au et al. [5] surveyed the permission systems of several popular smartphone operating systems and taxonomize them by the amount of control and information they provide users and the level of interactivity they require from users. Felt et al. [11] built a tool to determine the set of API calls that an application uses and then map those API calls to permissions. It generates the maximum set of permissions needed for an application and they compared them to the set of permissions actually requested.

However, these approaches are very hard to implement in practice. On the other hand, some researchers have dug into this area by constructing machine-learning-based researches. Wijesekera et al. [26] built a classifier to make privacy decisions on the user's behalf by detecting when context has changed and, when necessary, inferring privacy preferences based on the user's past decisions and behavior. It grants appropriate resource permission requests without further user intervention, denies inappropriate requests, and only prompts the user when the system is uncertain of the user's preferences. Li et al. [17] introduced Significant Permission IDentification (SigPID), a malware detection system based on permission usage analysis to cope with the rapid increase in the number of Android malware. They used several levels of pruning by mining the permission data to identify the most significant permissions. Then, they constructed machine-learning-based classifiers to classify different families of malware and benign apps.

Even so, users often do not fond of security software that frequently scan their devices. Therefore, Zhu et al. [32] introduced the techniques to automatically detect the potential security risk for each mobile App by exploiting the re-

quested permissions. Then, they designed a mobile App recommendation system with privacy and security awareness which can provide App recommendations by considering both the Apps' popularity and the users' security preferences. However, these approaches do not take the identity assets that Apps collect. Privacy risk exists because of insecure data access. Therefore, in this work we map each permission requested by mobile Apps to several identity assets and build our own privacy risk score software.

The other category is about mobile App's 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. However, not every application has a privacy policy. For example, Dehling et al. [9] surveyed popular medical health Apps in Apple iTunes Store and Google Play to assess the quality of medical health App's privacy policies. They found out that of the 600 most commonly used apps, only 183 had privacy policies. Liu et al. [19] 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. With the lack of taking user's privacy into concern, some works provide guidelines for building software and privacy policies. Harris [14] issued recommendations for mobile application developers and the mobile industry to safeguard consumer's 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.

Researchers have also begun to explore techniques for mitigating digital privacy risk. Zaeem et al. [28, 20, 27] proposed a technique that parses privacy policies and automatically generating summaries. They used data mining models to analyze the text of privacy policies, train their model with 400 privacy policies, and answer 10 basic questions concerning the privacy and security of user data. O'Loughlin et al. [22] reviewed data security and privacy policies of 116 mobile apps for depression. They constructed a list of questions and answer them by reviewing privacy policies. They showed that only 4% of privacy policies of mobile Apps are acceptable. Harkous et al. [13] proposed an automated framework for privacy policy analysis (Polisis). They built it with a novel hierarchy of neural-network classifiers and trained their model with 130k privacy policies. They provided PriBot which is a program that can answer users questions related to those privacy polices they have. Within 700 participants, PriBot's top-3 answers is relevant to users for 89% of the test questions. Nevertheless, these works do not look up what sets of identity assets are being collected by those privacy policies. Our work not only map permissions but also privacy policies to identity assets.

The last category is about security and privacy frameworks for mobile Apps. People have proposed scoring framework on social media. Petkos et al. [23] proposed a privacy scoring framework for Online Social Network (OSN) users with respect to the information about them that is disclosed and that can be inferred by OSN service operators and third parties. It took into account user's personal preferences, different types of information, and inferred information. To

fight against malwares, many works have been published to address data leakage problem. Rao et al. [25] 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 identity assets leaked from popular Apps and by malwares. Enck et al. [10] proposed a malware detection system named TaintDroid. "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. 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. Gibler et al. [12] 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.

Indeed, breaches of personal sensitive information can lead to gigantic damage to uses. To understand why such significant data leakage has occurred, Zuo et al. [33] designed tools for obfuscation-resilient cloud API identification and string value analysis, and implemented them in a tool called LeakScope to identify the potential data leakage vulnerabilities from mobile apps based on how the cloud APIs are used. On the other hand, Agarwal et al. [4] proposed ProtectMyPrivacy (PMP), a crowd sourced recommendation engine, to analyze manual protection decisions, and use them to provide iOS App privacy recommendations. It detects access to private information and protects users by substituting anonymized data based on user decisions. However, all the above recommendation approaches do not take consideration of the potential identity assets collected by mobile Apps, which motivates our novel work with awareness of permissions and privacy policies, which actually covers first and second categories.

5 Conclusion

In this paper, we sought to understand the privacy risk of the set of Personally Identifiable Information (PII), or identity assets, collected, used and shared by mobile applications. Each mobile App has a set of data access permissions and a privacy policy. Therefore, we sought to estimate the privacy risk score of each mobile App by investigating the set of identity assets that each mobile App collects, according to its privacy policy and data access permissions.

Our approaches leveraged the identity assets collected from these mobile apps and cross-referenced these PII to a list of over 600 identity assets collected in the Identity Theft Assessment and Prediction (ITAP) project at The University of Texas at Austin. The ITAP project investigates theft and fraud user stories to assess how identity asset is monetized and the risk (likelihood) of respective identity assets to be stolen and/or fraudulently used. From these mobile apps,

our results indicate that 67% of the over 600 reference identity assets were being collected by our sample dataset of 100 Android apps.

In this work, we proposed two approaches to estimate the privacy risk score of each mobile App. First approach is called Basic Measurement. It utilized the intrinsic characteristics of each identity asset to calculate the privacy risk score of each identity asset. The second approach is called Dynamic Measurement. It utilized two parameters that resulted from UT CID probabilistic models and Bayesian inference tool to refine the original risk of exposure and value of monetary loss. Our comparison with other researchers' work showed that our approaches are promising.

This work was the first to study privacy policies and permissions of mobile apps in terms of the identity assets collected, used and shared. We further studied those identity assets in the context of a personal data reference model built by the UT CID Identity Ecosystem and ITAP projects. This research provided a program to generate privacy risk score of each open-source mobile App and gave an empirical study of 100 open-source mobile Apps in different categories.

Acknowledgments This work was in part funded by the Center for Identity's Strategic Partners. The complete list of Partners can be found at the following URL: https://identity.utexas.edu/strategic-partners.

References

1. Google play, https://play.google.com/store
2. Immuniweb® mobile app scanner, https://www.htbridge.com
3. Itap report 2019. Tech. rep., Center for Identity, University of Texas at Austin (2019)
4. Agarwal, Y., Hall, M.: Protectmyprivacy: Detecting and mitigating privacy leaks on ios devices using crowdsourcing. pp. 97–110 (06 2013). https://doi.org/10.1145/2462456.2464460
5. Au, K., Zhou, Y., Huang, Z., Gill, P., Lie, D.: Short paper: A look at smartphone permission models. Proceedings of the 1st ACM Workshop on Security and Privacy in Smartphones and Mobile Devices (10 2011). https://doi.org/10.1145/2046614.2046626
6. Chang, K.C., Zaeem, R.N., Barber, K.S.: 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. pp. 114–120. ACM (2018)
7. Chang, K.C., Zaeem, R.N., Barber, K.S.: Internet of things: Securing the identity by analyzing ecosystem models of devices and organizations. In: 2018 AAAI Spring Symposium Series (2018)
8. Chen, C.J., Zaeem, R.N., Barber, K.S.: Statistical analysis of identity risk of exposure and cost using the ecosystem of identity attributes. In: 2019 European Intelligence and Security Informatics Conference (EISIC). pp. 32–39. IEEE (2019)
9. Dehling, T., Sunyaev, A., Taylor, P.L., Mandl, K.D.: Availability and quality of mobile health app privacy policies. Journal of the American Medical Informatics Association **22**(e1), e28–e33 (08 2014). https://doi.org/10.1136/amiajnl-2013-002605, https://doi.org/10.1136/amiajnl-2013-002605

10. Enck, W., Gilbert, P., Han, S., Tendulkar, V., Chun, B.G., Cox, L.P., Jung, J., McDaniel, P., Sheth, A.N.: Taintdroid: An information-flow tracking system for realtime privacy monitoring on smartphones. ACM Trans. Comput. Syst. **32**(2), 5:1–5:29 (Jun 2014). https://doi.org/10.1145/2619091, http://doi.acm.org/10.1145/2619091
11. Felt, A.P., Chin, E., Hanna, S., Song, D., Wagner, D.: Android permissions demystified. In: Proceedings of the 18th ACM Conference on Computer and Communications Security. p. 627–638. CCS '11, Association for Computing Machinery, New York, NY, USA (2011). https://doi.org/10.1145/2046707.2046779, https://doi.org/10.1145/2046707.2046779
12. Gibler, C., Crussell, J., Erickson, J., Chen, H.: Androidleaks: Automatically detecting potential privacy leaks in android applications on a large scale. pp. 291–307 (06 2012). https://doi.org/10.1007/978-3-642-30921-2_17
13. Harkous, H., Fawaz, K., Lebret, R., Schaub, F., Shin, K.G., Aberer, K.: Polisis: Automated analysis and presentation of privacy policies using deep learning. In: 27th USENIX Security Symposium (USENIX Security 18). pp. 531–548. USENIX Association, Baltimore, MD (Aug 2018), https://www.usenix.org/conference/usenixsecurity18/presentation/harkous
14. Harris, K.D.: Privacy on the go. Tech. rep., California Department of Justice (2013)
15. Hart, K.: Privacy policies are read by an aging few. Tech. rep. (2019)
16. Hornyack, P., Han, S., Jung, J., Schechter, S., Wetherall, D.: 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. pp. 639–652. CCS '11, ACM, New York, NY, USA (2011). https://doi.org/10.1145/2046707.2046780, http://doi.acm.org/10.1145/2046707.2046780
17. Li, J., Sun, L., Yan, Q., Li, Z., Srisa-an, W., Ye, H.: Significant permission identification for machine-learning-based android malware detection. IEEE Transactions on Industrial Informatics **14**(7), 3216–3225 (2018)
18. Liau, D., Zaeem, R.N., Barber, K.S.: Evaluation framework for future privacy protection systems: A dynamic identity ecosystem approach. In: 2019 17th International Conference on Privacy, Security and Trust (PST). pp. 1–3. IEEE (2019)
19. Liu, C., Arnett, K.P.: An examination of privacy policies in fortune 500 web sites. American Journal of Business **17**(1), 13–22 (2002). https://doi.org/10.1108/19355181200200001, https://doi.org/10.1108/19355181200200001
20. Nokhbeh Zaeem, R., Barber, K.S.: A study of web privacy policies across industries. Journal of Information Privacy and Security **13**(4), 169–185 (2017)
21. Nokhbeh Zaeem, R., Budalakoti, S., Barber, K.S., Rasheed, M., Bajaj, C.: Predicting and explaining identity risk, exposure and cost using the ecosystem of identity attributes. In: 2016 IEEE International Carnahan Conference on Security Technology (ICCST). pp. 1–8. IEEE (2016)
22. O'Loughlin, K., Neary, M., Adkins, E.C., Schueller, S.M.: Reviewing the data security and privacy policies of mobile apps for depression. Internet Interventions **15**, 110 – 115 (2019). https://doi.org/https://doi.org/10.1016/j.invent.2018.12.001, http://www.sciencedirect.com/science/article/pii/S2214782918300460
23. Petkos, G., Papadopoulos, S., Kompatsiaris, Y.: Pscore: A framework for enhancing privacy awareness in online social networks. In: 2015 10th International Conference on Availability, Reliability and Security. pp. 592–600 (2015)

24. Rana, R., Zaeem, R.N., Barber, K.S.: Us-centric vs. international personally identifiable information: A comparison using the ut cid identity ecosystem. In: 2018 International Carnahan Conference on Security Technology (ICCST). pp. 1–5. IEEE (2018)
25. Raoa, A., Kakhkib, A.M., Razaghpanahe, A., Tangc, A., Wangd, S., Sherryc, J., Gille, P., Krishnamurthyd, A., Legouta, A., Misloveb, A., Choffnesb, D.: Using the middle to meddle with mobile. Tech. rep., Northeastern University (2013)
26. Wijesekera, P., Baokar, A., Tsai, L., Reardon, J., Egelman, S., Wagner, D., Beznosov, K.: The feasibility of dynamically granted permissions: Aligning mobile privacy with user preferences. In: 2017 IEEE Symposium on Security and Privacy (SP). pp. 1077–1093 (2017)
27. Zaeem, R.N., Barber, K.S.: The effect of the gdpr on privacy policies: Recent progress and future promise. ACM Transactions on Management of Information Systems (2020), to Appear
28. Zaeem, R.N., German, R.L., Barber, K.S.: Privacycheck: Automatic summarization of privacy policies using data mining. ACM Trans. Internet Technol. **18**(4) (Aug 2018). https://doi.org/10.1145/3127519, https://doi.org/10.1145/3127519
29. Zaeem, R.N., Manoharan, M., Barber, K.S.: Risk kit: Highlighting vulnerable identity assets for specific age groups. In: 2016 European Intelligence and Security Informatics Conference (EISIC). pp. 32–38. IEEE (2016)
30. Zaeem, R.N., Manoharan, M., Yang, Y., Barber, K.S.: Modeling and analysis of identity threat behaviors through text mining of identity theft stories. Computers & Security **65**, 50–63 (2017)
31. Zaiss, J., Nokhbeh Zaeem, R., Barber, K.S.: Identity threat assessment and prediction. Journal of Consumer Affairs **53**(1), 58–70 (2019). https://doi.org/10.1111/joca.12191, https://onlinelibrary.wiley.com/doi/abs/10.1111/joca.12191
32. Zhu, H., Xiong, H., Ge, Y., Chen, E.: Mobile app recommendations with security and privacy awareness. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 951–960. KDD '14, Association for Computing Machinery, New York, NY, USA (2014). https://doi.org/10.1145/2623330.2623705, https://doi.org/10.1145/2623330.2623705
33. Zuo, C., Lin, Z., Zhang, Y.: Why does your data leak? uncovering the data leakage in cloud from mobile apps. In: 2019 IEEE Symposium on Security and Privacy (SP). pp. 1296–1310 (2019)

WWW.IDENTITY.UTEXAS.EDU