Risk assessment model of compromising personal data on mobile devices

Dmitry Izergin^{*}, and *Michael Ereemeev*

MIREA – Russian Technological University, 119454, Moscow, Russia

Abstract. Development of the information space to an avalanche-like increase in the volume of mobile data on the Internet. The generated digital portraits of users are becoming one of the main products for sale. The high quality of user digital portraits and their number is achieved through the use of intelligent data processing methods and the presence of large data sets. The volume of data processed by mobile devices and the number of modern services that collect various types of information make the issue of ensuring the confidentiality of user information the most important. Existing security mechanisms for mobile operating systems, as a rule, are aimed at neutralizing harmful effects and do not ensure the safety of personal data from legitimate services. The article proposes a model for assessing the risks of compromising personal data on mobile devices based on the correlation analysis of public information about service developers in order to detect the possibility of aggregating data from various sources.

1 Introduction

In the Russian Federation, the legal regulation of issues related to personal data is carried out by the Federal Law of July 27, 2006 No. 152-FZ "On Personal Data". According to article 3, clause 1 of the Federal Law-152, personal data include any information relating directly or indirectly to a specific or determined physical person - the subject personally [1]. This law is aimed at ensuring the protection of personal information (hereinafter PII, English Personal Information) relating to an identifiable person. In [2], the author classifies PII according to Figure 1.

^{*} Corresponding author: izergin@mirea.ru

Fig. 1. Classification of personal data.

The most important data include the “basic” category, which allows a person to be identified with a high degree of probability (Full name, passport data, date and place of birth, place of registration and place of actual residence). Data from the “Additional” category does not allow for personal identification, however, if there is a set of data or together with basic data, the probability of a determination error is minimized (information about education, social status, profession, phone number, insurance certificate number, etc.). Special personal data may contain information about ip - address, accounts, etc.

Biometric personal data includes information about the physiological and biological characteristics of a person. In the context of the development of modern technologies, based on a voice file or an image, it is possible to identify the user with high accuracy, which can serve as a basis for the recognition of media files as personal data. Additional, special and biometric data types (on the basis of which it is impossible to unambiguously determine their belonging to a specific person) will hereinafter be referred to as information correlated with a specific person (hereinafter ICP, English Information Correlation with the Person). As an example, the following data types can be referred to PII:

- Information about accounts on the device, containing name, home or work address, passport data with a photo or other linked documents;
- Fingerprint, iris information;
- Information about bank cards.

The data from the ICP category may include: phone number, voice messages, photo-video data, geolocation data, IP address, accounts in applications, health indicators, education, work, etc.

Thus, the amount and variety of processed data on mobile devices makes the issue of ensuring the safety of user information critically important. However, the existing protection mechanisms in modern mobile operating systems do not have functionality against the collection of personal data by legitimate services. Figure 2 shows the main types of sources used to compromise the identity of the device owner.

Fig. 2. Sources of user data leaks from mobile devices.

These types of sources include:

- operating system services and services installed by device manufacturers;
- browsers that collect and store data about visitors, as well as web resources that store the user's electronic resources (for example, tokens or cookies);
- legitimate applications that process personal data;
- software.

The study work analyses examine the sources of leaks from various types of applications running on mobile devices.

In works [3-7], studies were carried out to identify channels of data leaks. The analysis made it possible to identify the advantages of using the approaches. The article presents a model of risky data compromises on mobile systems, taking into account a set of applications. The relevance of the proposed model is due to the disclosure of the presented data of the device owner by services that aggregate data with various applications. The novelty of the model lies in taking into account the results of data analysis on the application device in order to detect the possibility of data aggregation from various sources.

2 Permission mechanisms

Modern mobile operating systems have various information security systems [8]. The permission mechanism is aimed at delimiting access to user data from installed applications and is designed to solve two main tasks:

- determination of the degree of confidentiality of the processed data (access levels);
- differentiation of access to hardware and software resources of the device based on user permission.

Figure 3 shows examples of screens for permission settings in iOS and Android [9,10]. In this article, the area of research is limited to studying the issue of countering personal data compromise for the Android platform.

Fig. 3. Permission engine settings screens in iOS and Android.

The Android operating system classifies the levels of access to functions according to the degree of criticality of the data being processed [9]. Access levels are classified into:

- normal - privileges that allow the application to access OS functions. Applications do not have the ability to compromise the user based on the level of access obtained (for example: setting the time zone, starting vibration, accessing information about the state of the network, etc.). No user permission required;
- dangerous - privileges that provide access to data by which it is possible to compromise the user's identity (for example: access to the phone book, location determination, message list, etc.). User permission is required;
- signatureOrSystem - privileges intended for applications located in the / system / priv-app directory (for example: uninstalling applications, installing applications, disabling application components, configuring WiFi, etc.).
- signature - privileges are available to applications signed with the same certificate as the application / firmware (examples: granting applications administrator rights, setting up the phone, mounting / unmounting the file system, getting a list of linked device accounts, managing applications).

Application developers interact with a limited list of possible system calls of the access level categories "normal" and "dangerous". The list of permissions for the "dangerous" category is shown in Figure 4.

Fig. 4. List of the main types of permissions in the category "dangerous".

Currently, there are thirty permissions that restrict access to PII (personally identifiable information) and ICP (Information Correlation with the Person), however, not all ICPs are protected by the privilege mechanism. Table 1 shows an example of permissions that do not require user consent to processing and possible vectors of compromise of the device owner.

Table 1. List of permissions leading to a PPII compromise.

Permissions	Description
ACCESS_NETWORK_STATE, CHANGE_NETWORK_STATE	Allows applications to access connection information (DHCP, DNS, Proxy, etc.). It is possible to define the egress nodes of the user's traffic.
ACCESS_WIFI_STATE, CHANGE_WIFI_STATE	Allows applications to access information about Wi-Fi networks. Obtaining information about the name of the WiFi network and mac - address in order to determine the location of the user.
BLUETOOTH_ADMIN	Allows applications to discover and pair Bluetooth devices. This functionality can be directed to scanning the environment in order to determine the user's location.

Assessing the potential for compromising PII requires monitoring application access to the Internet. Permission "INTERNET" does not apply to the category "dangerous", which is unacceptable when analyzing the possibility of user data leakage and must be taken into account when creating a model. It's worth noting that the permission mechanism is not an indicator of data leakage for the device owner. The inability to use applications or the inconvenience caused by restricting access, in most cases leads to a situation where the user allows access to the required functions of the device, which leads to the compromise of PII and ICP. To increase the level of information content and the quality of the granularity of the privilege mechanism, it is required to create new methods for protecting data, including from legitimate services.

3 Permission mechanisms

Modern mechanisms for ensuring the security of the functioning of a mobile device form an assessment of the state of the system based on the analysis of individual applications. This approach does not provide protection against the formation of a digital portrait of the user, since the ability to combine permissions from different applications in order to gain access to device resources is not taken into account. In order to improve the accuracy of risk indicators, a model is proposed that takes into account a set of applications. The proposed model is aimed at detecting channels that contribute to the construction of a digital portrait of a user by various services in order to aggregate data. An example of a scheme for constructing a digital portrait of a user is shown in Figure 5.

Fig. 5. The scheme for constructing a digital portrait of a user.

The existing mechanisms of privileges in mobile devices restrict access to information related directly or indirectly to an individual - the subject of personal data [9-10]. Applications that require a significant amount of the requested privileges may raise suspicion on the part of the user. In the case of using information from various services with different privileges, a high-quality solution to the problem is possible without attracting the attention of the device owner.

A prime example of a company that is open about aggregating information from various sources is Facebook Inc. This company has services such as Facebook, Instagram, Messenger, WhatsApp, WorkplaceChat, the Facebook Gaming platform and the analytical service Facebook Analytics. The analysis of data from various platforms allows you to form a high-quality profile of the device owner, which can be used to manipulate a person in a commercial and political environment. The risk of spreading a digital portrait of a user among interested companies is quite high. As an example, it is possible to cite the case of the transfer of this kind of information by Facebook Inc to Apple, Microsoft and Samsung corporations [11]. Accordingly, the use of knowledge about the connectivity between companies and services significantly increases the reliability of the risk assessment model for compromising user data.

4 Way to connect app developers

Various mobile app stores provide information on a set of services owned by one developer. The formation of a knowledge base of links between services and companies can be carried out on the basis of open information from trading floors. Figures 6 and 7 show Facebook applications in the official Google Play and AppStore stores.

Fig. 6. Facebook services on Google Play.

Fig. 7. Facebook services in the AppStore.

Based on the selected information, it is possible to create many applications with the same owner of the service (1):

$$D_{org} = \{I_{app_1}, I_{app_2} \dots I_{app_n}\}, \tag{1}$$

where I_{app_n} - is the application identifier

However, this method is superficial and does not take into account possible connections between different service owners. In the above example with Facebook Inc. there is no Instagram app owned by a subsidiary of Facebook. In order to increase the likelihood of service binding, it is necessary to additionally analyze the information on the application pages. Automated binding of service developers is carried out based on the extraction of features from open sources. The distinguishing features include:

- developer contact information (2):

$$L_{app} = \{L_{email}, L_{address}, L_{website}\}, \tag{2}$$

where L_{email} – set of postal addresses; $L_{address}$, – set of legal addresses of the organization, or actual location; $L_{website}$ – a set of links to the owner.

- user agreement (3):

$$S_{app} = \{S_{key}, S_{link}, L_{app}\}, \tag{3}$$

where S_{key} – a set of keywords; S_{link} – a set of links to web resources; L_{app} – developer contact information.

Figure 8 shows the linking scheme for Instagram service developers and Facebook based on public information.

Fig. 8. Linking service developers based on public information.

The formation of intersections $L_{address}$ is based on the construction of a digital fingerprint of the data of each address, in which close fingerprints correspond to similar data (perceptual hashing). The estimation of distance / similarity measures of strings in the vector representation P_f is based on the SimHash algorithm. The collision probability F depends on the similarity of the arguments x, y (4):

$$P_F\{h(x) = h(y)\} = (x, y), \quad (4)$$

where $h(x) = \text{sign}(R_x), h(y) = \text{sign}(R_y)$ with piecewise operation sign , – a measure of similarity of objects x, y , taking a value in $[0,1]$. The components of d - dimensional vectors are a set of d values of the locally sensitive hash $h_i(x/y), i \in [d]$. The estimate of the likelihood $P^*\{h(x) = h(y)\}$ of similarity of company addresses is calculated as follows:

$$P\{h(x) = h(y)\} = \frac{\sum_{i=1}^d \{h_i(x) = h_i(y)\}}{d} = 1 - \text{dist}_{\text{ham}}(h(x), h(y)), \quad (5)$$

where $\text{dist}_{\text{Ham}}(h(x), h(y)) = \sum_{i=1}^d \{x_i \neq y_i\}$ Hamming distance between binary vectors $h(x)$ and $h(y)$.

The processing of the sets $L_{email}, L_{website}, S_{key}, S_{link}$ of various applications consisted of intersecting them and finding common values. Applying the proposed method of linking Facebook and Instagram will provide the following result:

$$S_{\text{linkInstagram}} \cap S_{\text{linkFacebook}} = \{\text{https://www.facebook.com}\}, \quad (6)$$

$$S_{\text{keyInstagram}} \cap S_{\text{keyFacebook}} = \{\text{Facebook, Instagram, Messenger, WhatsApp, Oculus}\}, \quad (7)$$

$$P\{h(L_{\text{appInstagram}}) = h(L_{\text{appFacebook}})\} = 0,99, \quad (8)$$

The result obtained makes it possible to take into account the degree of connectivity of service owners when calculating the risks of compromising personal data by a mobile device. The threshold value for the activation level of the binding procedure is set depending on the system requirements. The result of the given method represents a set of actual owners of services or subsidiaries:

$$H_{\text{corp}} = \{D_{\text{org}_1} \dots D_{\text{org}_n}\}, \quad (9)$$

where $D_{\text{gn}} = \{I_{\text{app}_1} \dots I_{\text{app}_n}\}$ – the declared owner of the service.

After the procedure for linking service owners, it is required to assess the possible quality of the digital portrait of the user, based on the transmitted data from various applications. When forming U_{org} , the presence of the same data types in the sets U_{PII_i} и U_{ICP_i} is possible, which is redundant when assessing the completeness of a digital portrait. A digital portrait of a user can be represented as:

$$U_{\text{org}} = \left\{ \sum_{i=1}^n |U_{\text{PII}_i}| - \sum_{i < j}^n |U_{\text{PII}_i} \cap U_{\text{PII}_j}|, \sum_{i=1}^n |U_{\text{ICP}_i}| - \sum_{i < j}^n |U_{\text{ICP}_i} \cap U_{\text{ICP}_j}| \right\}, \quad (10)$$

where U_{PII_i} – a lot of personal data processed by the application; U_{ICP_i} – a set of processed data, correlated with a specific person.

Let us denote the probability of user identification by applications A and B belonging to the same owner as $P(A)$ and $P(B)$, respectively. Since the likelihood of user identification by different applications are compatible and independent events, then in the case of data aggregation, the total probability of events can be represented as:

$$P_{\text{org}} = P(A) + P(B) - P(A) \times P(B), \quad (11)$$

Thus, the mathematical model for assessing the risks of compromising personal data on a mobile device can be presented in the following form:

$$S_{\text{dev}} = \begin{cases} f(T_{\text{sim}_i}) = \sum_{i=1}^n \text{Sim}_{\text{app}_i}(T_{\text{app}_i}, \text{Pattern}_{\text{category}}); \\ f(T_{U_{\text{org}}}) = \sum_{i=1}^n U_{\text{org}}; \\ f(T_{P_{\text{org}}}) = \sum_{i=1}^n P_{\text{org}}; \\ f \in H \end{cases}, \quad (12)$$

where f – decision rule (function); T_{app_i} – digital portrait of the application.

5 Conclusion

Currently, modern privacy systems do not have the ability to counteract the process of building a digital portrait of a user. The user is not allowed by privilege mechanism to determine the degree of threat to confidential data, and the issue of countering the collection and aggregation of information from various applications is unresolved. The risk of compromising Personally Identifiable (PII) and Personally Correlated (ICP) data across a suite of applications after various aggregation and post-processing processes can significantly improve the digital portrait of the device owner. The result of the work is aimed at increasing the level of protection of personal data and ensuring the ability to track the aggregation of information from various sources. A method for linking application developers based on the correlation of data extracted from security policies and contact

information is presented. The result of this work is a model of the functioning of the personal data leak detection system.

In further studies, it is planned to test the developed model of the functioning of the personal data leak detection system using various machine learning methods to identify the most optimal algorithm for determining risk indicators.

References

1. Federal'nyj zakon ot 27.07.2006 N 152-FZ (red. ot 30.12.2020) «O personal'nyh dannyh» (s izm. i dop., vstup. v silu s 01.03.2021)
2. N. Platonova, *Pravo i sovremennye gosudarstva*, **5**, 9 (2017)
3. N. Wongwiwatchai, P. Pongkham, K. Sripanidkulchai, *Computers & Security*, **99**, 102011 (2020)
4. J. Yang, C. Kim, M. Hassan Onik, 21st International Conference on Advanced Communication Technology, 425 (2019)
5. M. Hassan Onik, C. Kim, N. Lee, J. Yang, *Applied Sciences*, **9**, 3997 (2019)
6. A. Kumar Jha, Woo Jin Lee, *Journal of Systems and Software*, **137**, 550 (2018)
7. X. Zheng, Z. Cai, Y. Li, *IEEE Communications Magazine*, **56(9)**, 55 (2018)
8. D. Izergin, M. Ereemeev, Sh. Magomedov, S. Smirnov, *Rossijskij tehnologicheskij zhurnal*, **7(6)**, 44 (2019)
9. Permissions overview, URL: <https://developer.android.com/guide/topics/permissions> (accessed 20.04.2021)
10. Requesting Permission, URL: <https://developer.apple.com/design/human-interface-guidelines/ios/app-architecture/requesting-permission/> (accessed 20.04.2021)
11. Kak Facebook sobiraet i prodaet dannye pol'zovatelej, URL: https://www.tadviser.ru/index.php/Статья:Как_Facebook_собирает_и_продает_данные_пользователя (accessed 28.12.2020)
12. A. Skovoroda, D. Gamajunov, *PDM*, **36**, 84 (2017)
13. D. Rachkovskij, *Kibernetika i sistemnyj analiz*, **53(№1)**, 160 (2017)
14. L. Fritsch, N. Momen, In Open Identity Summit (OID). Lecture Notes in Informatics (LNI), **277** (2017)
15. J. Lopez, R. Rios, F. Bao, G. Wang, *Future Generation Computer Systems*, **75**, 46 (2017)
16. I. Chong, H. Ge, N. Li, R. Proctor, *Computer Security*, **78**, 143 (2018)
17. E. Ahmed, I. Yaqoob, I.A.T. Hashem, J. Shuja, M. Imran, N. Guizani, S.T. Bakhsh. *IEEE Commun. Mag*, **56**, 102 (2018)
18. K. Kenthapadi, I. Mironov, A. Thakurta, WWW '19 - Privacy-preserving Data Mining in IndustryTaipei, 1308 (2019)
19. Y. He, X. Yang, B. Hu, W.J. Wang, *Inf. Secur. Appl*, **46**, 259 (2019)
20. K. Ito, H. Hasegawa, Y. Yamaguchi, H.Y. Shimada. In Proceedings of the 2018 International Workshop on Security, 143 (2018)
21. M.M.H. Onik, N. Al-Zaben, J. Yang, N.Y. Lee, C.S. Kim, In Proceedings of the International Conference on Computing, Electronics & Communications Engineering 2018 (icCECE '18), 71 (2018)
22. A. Arora, S.K. Peddoju, V. Chouhan, A. Chaudhary, In Proceedings of the 24th Annual International Conference on Mobile Computing and Networking, 798 (2018)

23. A. Razaghpanah, R. Nithyanand, N. Vallina-Rodriguez, S. Sundaresan, M. Allman, C. Kreibich, P. Gill, A Global Study of the Mobile Tracking Ecosystem. In Proceedings of the 2018 Network and Distributed System Security Symposium, 18 (2018)
24. R. Baalous, R. Poet, In Proceedings of the 11th International Conference on Security of Information and Networks, 26 (2018)