How Dangerous Permissions are Described in Android Apps' Privacy Policies?

Rawan Baalous
School of Computing Science
University of Glasgow
Glasgow, United Kingdom
r.baalous.1@research.gla.ac.uk

Ronald Poet
School of Computing Science
University of Glasgow
Glasgow, United Kingdom
ron.poet@glasgow.ac.uk

ABSTRACT

Google requires Android apps which handle users' personal data such as photos and contacts information to post a privacy policy which describes comprehensively how the app collects, uses and shares users' information. Unfortunately, while knowing why the app wants to access specific users' information is considered very useful, permissions screen in Android does not provide such pieces of information. Accordingly, users reported their concerns about apps requiring permissions that seem to be not related to the apps' functions. To advance toward practical solutions that can assist users in protecting their privacy, a technique to automatically discover the rationales of dangerous permissions requested by Android apps, by extracting them from apps' privacy policies, could be a great advantage. However, before being able to do so, it is important to bridge the gap between technical terms used in Android permissions and natural language terminology in privacy policies. In this paper, we recorded the terminology used in Android apps' privacy policies which describe usage of dangerous permissions. The semi-automated approach employs NLP and IE techniques to map privacy policies' terminologies to Android dangerous permissions. The mapping links 128 information types to Android dangerous permissions. This mapping produces semantic information which can then be used to extract the rationales of dangerous permissions from apps' privacy policies.

CCS CONCEPTS

• **Security and privacy → Privacy protections**

KEYWORDS

privacy policy, dangerous permissions, apps, Android

1 INTRODUCTION

Many mobile apps collect users' personal information. In order to protect users' privacy, the Android security model requires apps to declare the permissions they need to access in the manifest file [1]. However, users reported their concerns about apps requiring permissions that seem to be not related to the apps' functions [2,3].

Google requires apps which handle sensitive or personal user information to post a privacy policy which comprehensively describes how the app collects, uses and shares users' information [4]. However, many users don't read privacy policies and ignore them. This is not because they don't care about their privacy, but because privacy policies are too long and some information is hidden in the text [5].

It is challenging for users to know how their data are being used by Android apps, thereby make it difficult for them to assess potential risks [6]. Providing users with rationales of requested permissions by extracting them from the apps' privacy policies can play an important role in addressing users' doubts and unanswered questions. Before being able to do so, we need first to bridge the gap between technical terms used in Android permissions and natural language terminology in privacy policies. In this work, we recorded privacy policies' terminologies which describe usage of dangerous permissions. We chose to address dangerous permissions only in Android for the following reasons: first, our study is motivated by users' privacy concerns, hence we only considered dangerous permissions that may affect users' privacy. Second, these are the permissions that the user has a control of in the Android device setting, by allowing or revoking them at any time.

The rest of this paper is structured as follows: in Section 2 we detail the method used to extract privacy policies' terminologies that are related to Android dangerous permissions. In Sections 3 we present the results and discuss them. Finally, we draw the conclusion in Section 4.

2 METHODOLOGY

In order to construct our data set of Android apps' privacy policies, the top 100 Android apps were chosen from Google play. The chosen apps covered all Google Play categories. Since some apps don't provide a privacy policy link on Google Play, and some apps share the same privacy policy document (i.e., Google apps that share the same Google privacy policy document) and we only consider English privacy policies, we ended up with 73 available unique English privacy policies as our data set. The process of choosing top Android apps from Google play market took place on December 7th 2017. The methodology used to map privacy policies' terminologies to Android dangerous permissions is illustrated in detail in the following sections.

2.1 Pre-Process the Privacy Policy

Since the HTML privacy policy's file contains meta tags, JavaScript, style, website navigation, etc., we need first to navigate through HTML tags and pull the privacy policy text out of it. To do so, we used Beautiful Soup Python Library [7]. After

SIN '18, September 10–12, 2018, Cardiff, United Kingdom
© 2018 Copyright is held by the owner/author(s).
ACM ISBN 978-1-4503-6608-3/18/09.
https://doi.org/10.1145/3264437.3264477

that, the privacy policy text was segmented into sentences. We used the Natural Language Toolkit [8] for sentence segmentation. Then, the privacy policy text was converted to lowercase.

2.2 Extract Data Collection, Usage, Sharing and Retaining Practices

In the second step, we want to extract sentences containing data collection, usage, sharing and retaining practices in privacy policy document. In order to do so, we extracted all sentences containing any of the most common used verbs in privacy policies which represent collecting, using, sharing or retaining of user's information. These common verbs are based on [9] and [10] studies. From the first study, we only included verbs that specify a collection, usage, sharing, or retaining data practice. Other verbs such as "notify" which is used to make users aware of organization's data practices were excluded. On the other hand, in the second study, the authors traced privacy policies' keywords which indicate that the action is to be classified as a collection, usage, sharing, or retention. Therefore, all their resulted verbs were included in our study.

2.3 Identify Information Types

In the third step, we automatically identified noun phrases from the extracted sentences. In order to identify noun phrases, we used TextBlob [11], a Python library for processing textual data. Using TextBlob, each sentence was split into tokens. Next, each token was assigned a part-of-speech (POS) tag, such as "VB" for verb, "NN" for noun and "JJ" for adjective, among others. Finally, noun phrases which match the patterns presented in Table 1 were automatically extracted. After having a set of identified noun phrases from each sentence, we manually reviewed each noun phrase. If the identified noun phrase is not an information type, then it is excluded.

Finally, we manually scanned the privacy policy to include missed information types that are collected, used, shared or retained according to the specified noun phrases' patterns. This is to improve coverage of results and to ensure reliable ground truth data as privacy policies may use different ways to express data handling practices.

Table 1: Noun phrases' patterns

Noun Phrase Pattern	Example
{NN}	Contacts
{JJ} + {NN}	Personal information
{NN} + {NN}	Contact list

2.4 Map Information Types to Android Dangerous Permissions

The resulting information types are then mapped into each dangerous permission (i.e., each information type can be mapped to one or more dangerous permission) using hypernym, synonym and meronym relationships. The final resulted mappings were reviewed by two domain experts: a lawyer and an Android developer. We only considered mappings where both experts agreed on the mapping.

3 RESULTS

The 73 Android apps' privacy policies consisted of a total of 128 information types that are matched to Android dangerous permissions. The most interesting characteristic of the mapping results is the wide variety of terminologies used by privacy policies for dangerous permissions. For example, privacy policies use the terms: "address book" and "device's phonebook" to describe that they are using the "Contacts" permissions. Table 2 presents the frequency of the extracted privacy policies' terminologies, which are related to Android dangerous permissions. The table only presents a subset of the all 128 privacy policies' terminologies.

Table 2: Frequency of privacy policies' terminologies

Privacy Policy's Terminology	Frequency
Personal information	74%
Phone number/s	34%
Location	30%
Location information	29%
Photo/s	27%
Contact information	26%
Personal data	23%
Telephone number/s	22%
Location data	16%
Address	15%

4 CONCLUSION

In this work, we have presented a semi-automated analysis of the terms used in Android apps' privacy policies, with the aim of finding hypernym, synonym and meronym concepts and establishing their relationships to Android dangerous permissions. The results of the analysis provided 128 privacy policies' terminologies that are matched to Android dangerous permissions. The findings of this work provide the ground truth data for future research in which the rationales of dangerous permissions will be automatically extracted from Android apps' privacy policies.

REFERENCES

[1] Li, Y., Chen, F., Li, T.J.J., Guo, Y., Huang, G., Fredrikson, M., Agarwal, Y. and Hong, J.I., (2017). PrivacyStreams: Enabling Transparency in Personal Data Processing for Mobile Apps. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 1(3), p.76.

[2] Manjunath, (2015). Permission while downloading ola app. [online] Available at:https://support.google.com/googleplay/forum/AAAA8CVOtD8aN_Uym1aFD0?hl=en [Accessed 29 Dec. 2017].

[3] Wilson, R.D., (2015). Excessive permissions. [online] Available at: https://support.google.com/googleplay/forum/AAAA8CVOtD8wBx9IF2zwTs?hl=en [Accessed 29 Dec. 2017].

[4] Google, (2017). Privacy, Security, and Deception. [online] Available at: https://play.google.com/about/privacy-security-deception/personal-sensitive/ [Accessed 28 Dec. 2017].

[5] Costante, E., Den Hartog, J. and Petkovic, M., (2011), September. On-line trust perception: What really matters. In Socio-Technical Aspects in Security and Trust (STAST), 2011 1st Workshop on (pp. 52-59). IEEE.

[6] Gibler, C., Crussell, J., Erickson, J. and Chen, H., (2012), June. AndroidLeaks: automatically detecting potential privacy leaks in android applications on a large scale. In International Conference on Trust and Trustworthy Computing (pp. 291-307). Springer, Berlin, Heidelberg.

[7] Beautiful Soup. (2004). [online] https://www.crummy.com/software/BeautifulSoup/

[8] NLTK. (2005). [online] http://www.nltk.org/

[9] Anton, A.I. and Earp, J.B., (2004). A requirements taxonomy for reducing web site privacy vulnerabilities. Requirements Engineering, 9(3), pp.169-185.

[10] Breaux, T.D., Hibshi, H. and Rao, A., (2014). Eddy, a formal language for specifying and analyzing data flow specifications for conflicting privacy requirements. Requirements Engineering, 19(3), pp.281-307.

[11] TextBlob. (2013). [online] https://github.com/sloria/textblob