I Don't Know Why You Need My Data: A Case Study of Popular Social Media Privacy Policies

Elizabeth Miller, Md Rashedur Rahman, Moinul Hossain, Aisha Ali-Gombe
Towson University
{emille51, mrahma4}@students.towson.edu
{mhossain, aaligombe}@towson.edu

Abstract

Data privacy, a critical human right, is gaining importance as new technologies are developed, and the old ones evolve. In mobile platforms such as Android, data privacy regulations require developers to communicate data access requests using privacy policy statements (PPS). This case study cross-examines the privacy policy in popular social media (SM) apps — Facebook and Twitter — constituting 314 candidate statements for features of language ambiguity, sensitive data requests, and whether the statements tally with the data requests made in the Manifest file. Subsequently, we conduct a comparative analysis between the PPS of these two apps to examine trends that may constitute a threat to user data privacy.

CCS Concepts

• **Security and privacy → Social aspects of security and privacy**.

Keywords

Privacy Policy, Facebook, Twitter, Social Media, Privacy Policy Analysis

ACM Reference Format:
Elizabeth Miller, Md Rashedur Rahman, Moinul Hossain, Aisha Ali-Gombe. 2022. I Don't Know Why You Need My Data: A Case Study of Popular Social Media Privacy Policies. In *Proceedings of the Twelveth ACM Conference on Data and Application Security and Privacy (CODASPY '22), April 24–27, 2022, Baltimore, MD, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3508398.3519359

1 Introduction

As of 2021, there are approximately 2.8 billion Android device users in the world with 2.56 million apps available to download via the Google Play store and the most popular of these applications are social media apps. In the United States, 82% of the population has a social networking profile [3]. With a significant percentage of people using social media applications, user privacy has become an ever-increasing concern [1]. Regulations — e.g., the European General Data Protection Regulation (GDPR)[4], and the California Consumer Privacy Act (CCPA)[2] — have been put in place to address these privacy concerns and guarantee that users provide informed consent to these social media apps requesting the usage of their data. These regulations mandate that a data request must be made unambiguous. More importantly, the type of data, the reason for the request, and in some cases, the purpose limitation must be stated and approved by the user ahead of time. The permission model is a dedicated system in the Android framework that ensures users give explicit access to their personal or device data. Unfortunately, in its current design, this model does not address why the data is requested, its destination, and with whom it would be shared. While an improved version of this model designed to address the permission intent [5] has been proposed in the literature, it is not yet adopted into the Android system. Thus, for developers to comply with the stated regulations, they often leverage the combination of this permission model in conjunction with a privacy policies statement (PPS). However, given the lack of standardization in PPS, many developers have resorted to exploiting these contracts using vague and ambiguous legal jargon to request data access and declare reason and sharing limitations.

Thus, the fundamental goal of our research is to determine how comprehensible various social media privacy policies are. To evaluate this, we investigate the vagueness and language ambiguity of PPS in Facebook and Twitter apps. Our study examines: (1) whether these apps clearly and unambiguously ask for user permission in the PPS and the level of sensitivity of requested data, (2) whether the data requested in the PPS tallies with the explicit data requests made during execution, (3) a comparative analysis of the PPS of these two apps to identify trends in vagueness and sensitivity.

2 Methodology

In this research, we leveraged case study methodology to examine the PPS and permission list of Facebook and Twitter directly from Google play. Using a four-step process, we manually examine every statement in the PPS for our target apps.

(I) Language Extraction - The first step is to manually read through the PPS for Facebook and Twitter and look for user data request statements called the candidate statements. We defined candidate statements as statements that contain three primary elements: i) a focused verb representing the data access action, e.g., transfer, obtain, etc. ii) a noun that identifies the type of data being accessed, and iii) a description of how the app will use the specified data type. An example of a data request statement is "We use your location data to recommend restaurants near you." In this example, the requesting verb is *use* which shows that the app is accessing user information. The data type in this example is *location data*, while *to recommend restaurants near you* describes how the app plans to use the data. All candidate statements from each PPS are manually extracted, deconstructed, and recorded in the Results_Table 1 using this verb-data-purpose mapping technique.

(II) Data Clustering - We use data clustering to organize the Results_Table from task 1 and group synonymous data types. For

CODASPY '22, April 24–27, 2022, Baltimore, MD, USA

ACM ISBN 978-1-4503-9220-4/22/04.
https://doi.org/10.1145/3508398.3519359

Focused Verb	Data Requested	Permission Name	Declared Intent (Usage)	Is Vague	Is Sensitive	Permission Asked
collect	photos	Storage	To create content	No	Yes	Yes
use	audio data	Record Audio	N/A	Yes- usage ambiguity	No	No
collect	credit card info	N/A	To make a purchase	No	Yes	No

Table 1: Comparative Analysis of the Case study

instance, data types such as "microphone data" and "audio recordings" are simplified and categorized as "audio data." After that, each data type is examined to determine whether it is sensitive according to the legal definition of sensitive data. Sensitive data refers to "information that is protected against unwarranted disclosure, to include Personally Identifiable Information (PII), Protected Health Information (PHI) or other private/confidential data, as specifically determined by the State." This includes any information that can be used to identify an individual or is linked to an individual.

(III) Data Mapping - It is important to note that not all PIIs have corresponding permissions in the Android permission model (e.g., SSN), and some vital user and device data do not fall into the legal definition of sensitive data (e.g., SMS). Thus, to establish that a user will receive a runtime explicit request for sensitive data, we leverage data-permission mapping to identify additional sensitive data groups in our Results_Table backed by permission. We conduct a side-by-side comparison of the Android permission list for each app provided in Google Play with the data in our Results_Table. We recorded the *Permission Name* as a new column and marked the *Permission Asked* column as yes in the table if its corresponding data is in PPS and the permission list. Otherwise, we record the *Permission Name*, and the *Permission Asked* as no. For PIIs not backed by permission, we marked the *Permission Name* as N/A, and *Permission Asked* as no.

(IV) Ambiguity Analysis - Finally, we analyze each candidate statement for features of ambiguity or vagueness. One of two subcategories determines the vagueness of a particular privacy statement: 1) data ambiguity, and 2) usage ambiguity. In the first case, the ambiguity is attributed to the data type when it is unclear what information is requested. For example, the statement "We access your information to provide our services to you" would be identified as *vague-requesting data* because the requesting data type (noun) in this example is "your information," which does not clearly specify what data the app is accessing. In the second case, if the intent for data access is unstated or unclear, a statement is marked as *vague-potential usage.* A statement such as "We collect your audio data." is an example of usage ambiguity since there is no explicit declaration of why the data is requested.

3 Data Collection and Analysis

3.1 Data Collection

Using our 4-step process methodology, we populated the Results _Table as shown in Table 1 with the data collected from the deconstructed statements collected from the PPS of Facebook and Twitter. Our data collection resulted in a seven-column table with 314 entries representing the number of candidate statements examined. The columns describe the focused verb, type of data requested, whether the requesting data is sensitive, whether the app declared permission for the data in the manifest, associated permission name, declared intent (usage), and whether the statement is vague.

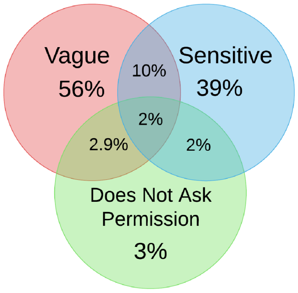

Figure 1: PPS Data Analysis Result for Facebook

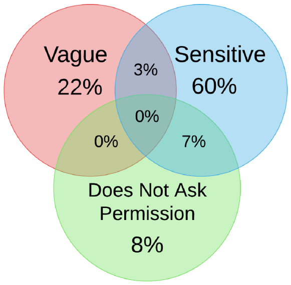

Figure 2: PPS Data Analysis Result for Twitter

3.2 Data Analysis

Upon completing all the data collection for our target apps, we analyzed the data entries for each app separately. This process allows us to calculate the percentages of all vague and sensitive statements as well as statements that do not ask for user permission. It also allows us to identify any existent overlap between these three categories, including the intersection of sensitive and vague statements, sensitive statements that do not ask for permission, and statements that fit all three categories simultaneously.

Facebook - In total, 207 candidate statements were extracted and analyzed from Facebook. Of these privacy policy statements we found that:

- 56.04% of statements were flagged as vague.
  - 60% of vague statements were flagged as vague due to the possible usage.
  - 40% of vague statements were flagged as vague due to the data type.
- 38.7% of statements involved sensitive data types (PII and none-PII).
- 3.4% of statements did not ask for permission via the Android permission model.

We then analyzed the overlap of PPS data analysis results for Facebook as shown in Figure 1. We found that:

- 10% of statements were both vague and strictly involved sensitive user data.
- 2% of statements did not ask for user permission and involved sensitive user data.

- 2.9% of statements did not ask for user permission and were vague.
- 2% of statements fit all 3 categories- vague, sensitive, and did not ask for user permission.

Twitter - For Twitter, a total of 107 privacy statements were reviewed. From these statements we found that:

- 22.43% of statements were flagged as vague.
  - 54% of vague statements were flagged as vague due to the possible usage.
  - 46% of vague statements were flagged as vague due to the data type.
- 60% of statements involved sensitive data types (PII and none-PII).
- 8.4% of statements did not ask for permission via the Android permission model.

Similar to Facebook, we analyze the overlap of all three categories, as shown in the center of Figure 2. The results indicate that:

- 3% of statements were both vague and strictly involved sensitive user data.
- 7% of statements did not ask for user permission and involved sensitive user data.
- 0% of statements did not ask for user permission and were vague.
- 0% of statements fit all 3 categories- vague, sensitive, and did not ask for user permission.

Unknown Sensitivity and Permissions - It is important to note that for statements flagged as "vague and sensitive" or "vague and does not ask permission," we only included those statements for which the sensitivity and permissions could be confirmed were included. For instance, in a statement such as "We collect your information to personalize our services for you," the identified data type is "your information." In this case, we do not know for certain what information is being collected. Therefore, we cannot determine if this information is sensitive or if the app is asking permission for this data. Such statements are flagged as "vague" but cannot be flagged as "sensitive" or "does not ask permission ." Thus, we put this statement in a separate category - *Unknown Sensitivity and Permissions.*

3.3 Comparative Analysis

Examining the results of Facebook and Twitter side-by-side, we found that both apps requested a large number of sensitive data such as (payment information, IP address, location, account information/-password, contact information). The results indicate that Twitter requested sensitive data in 64 statements compared to Facebook's 80 statements. More so, Facebook PPS has a higher percentage of vague statements (50%) than Twitter (22%). These vague statements correspond to more than half of all the candidate statements examined for Facebook (≈103), with more than 42 data ambiguity statements. In contrast, Twitter recorded 22 vague statements, with about 10 data ambiguity statements. It is also important to note that, for both the two apps, usage ambiguity takes the higher percentages (60% and 54%), thus indicative that apps seldom provide reasons for data requests to the user. Additionally, comparing the number of vague statements requesting sensitive data, we found that Facebook (10%) is again higher than Twitter (3%). This percentage shows that more than 20 of all the candidate's statements examined for Facebook requested both sensitive data and are vague in specifying why the data is asked (ambiguity of usage). On the other hand, Twitter has seven statements that did not ask for user permission and involved sensitive data compared to Facebook's four statements. Another notable distinction between these two apps is that 2% of Facebook's PPS intersected in the three categories. Twitter's PPS, on the other hand, did not contain any such statements. Finally, we found that about 22% of all the candidate statements analyzed for Facebook fall into the unknown sensitivity and permission category. For Twitter, roughly about 13% falls into this category. As a result, the percentages of "sensitive and vague" and "vague, and do not ask permission" statements are equivalent to or higher than the percentages we reported.

Thus, our findings from this study indicate that Facebook has more ambiguous statements that lack clarity both in terms of data requests and usage in its privacy policy statement. The candidate statements, especially those that fall into two and three-category overlap, need to be carefully reviewed by the developers.

Future Work - This study is limited to exploring the PPS of only two apps. Although these apps are the two most popular SM apps, they are not good representatives of the population. Thus, we plan to extend this research to include more SM applications as part of future work. In addition, we plan to manually generate a large corpus of deconstructed PPS that will enable us to leverage NLP for the automated detection of ambiguity and policy vagueness.

4 Conclusion

This study explored the level of ambiguity, sensitivity, and whether built-in PPS tally with the runtime permission requests for user data in two SM apps. Our results showed a significant portion of all the analyzed PPS statements for Twitter and Facebook requests for "very sensitive" user data. We also demonstrated that more than half of all the PPS analyzed, especially for Facebook, have some form of data or usage ambiguity. Of those analyzed statements, a substantial percentage falls into the intersection of vagueness (usage ambiguity) and sensitive data, thus indicative that users are not provided with clear and informed consent, thereby posing a potential threat to their privacy. Finally, an important finding in this study is that both apps have a substantial number of statements that fall into the unknown sensitivity and permissions category, which warrants further investigation.

5 Acknowledgements

This work is supported by the NSF Grant # 1850054 and Towson University OSPR.

References

[1] Ghazaleh Beigi and Huan Liu. 2020. A survey on privacy in social media: identification, mitigation, and applications. *ACM Transactions on Data Science* 1, 1 (2020), 1–38.

[2] CCPA. [n.d.]. California Consumer Privacy Act (CCPA). https://oag.ca.gov/privacy/ccpa. (Accessed on May 29, 2021).

[3] Statista Research Department. 2021. Percentage of U.S. population who currently use any social media from 2008 to 2021. https://www.statista.com/statistics/273476/percentage-of-us-population-with-a-social-network-profile/. (Accessed on December 15, 2021).

[4] GDPR. [n.d.]. General Data Protection Regulation- GDPR. https://gdpr-info.eu/. (Accessed on May 29, 2021).

[5] Md Rashedur Rahman, Elizabeth Miller, Moinul Hossain, and Aisha Ali-Gombe. 2022. Intent-Aware Permission Architecture: A Model for Rethinking Informed Consent for Android Apps. *arXiv preprint arXiv:2202.06995* (2022).