Thomas Brüggemann

Master Thesis
im Fach Information Systems

Automated Information Privacy Risk Assessment of Android Health Applications

Themensteller: Prof. Dr. Ali Sunyaev

Vorgelegt in der Masterprüfung
im Studiengang Information Systems
der Wirtschafts- und Sozialwissenschaftlichen Fakultät
der Universität zu Köln

Köln, September 2016

Contents

Index of Abbreviations

mHealth	Mobile Health

List of Tables

1. Introduction

1.1 Problem Statement

The market for mobile phone and tablet applications (apps) has grown extensively since recent years.[1] It is increasingly easier for companies or even single developers to create unique apps that reach millions of users around the planet via digital app stores. This market growth affected mobile health (mHealth) apps as well. More and more mHealth apps are available that support the users in resolving their health-related issues and that try to remedy health-related information deficiencies.

But receiving personal health-related information yields information privacy risks to users. Users are asked to expose personal health-related information, e.g. information on disease symptoms or medical appointments in order to receive a tailored app that fits their needs.[2] It remains however unclear how and where the vulnerable user information is sent, processed and stored.[3]

The information about these privacy related practices of app providers and their offered apps should be stated in the privacy policy document provided by the app provider.[4] Processing these privacy policies requires a higher level of education and time to read through large bodies of text, in order to find the relevant information. Additionally, the important information is hidden in legal language or is insufficiently addressed, if at all.[5] Aside from data usage beyond the control of the users, it is also challenging to assess what kind of private information an app asks for, prior to the app usage. Users have to download the apps of interest and try them out, before it becomes clear what health-related information is processed by the app and in which way. This leads to low comparability between apps. When users are looking for specific functionality in an mHealth app, it is challenging to find the app that offers the desired functionality at an acceptable information privacy risk. Even if users would pursue the task of finding and comparing mHealth

---

1 See for this and the following sentence Enck et al. (2011), p. 1.

2 See Chen et al. (2012), p. 2.

3 See He et al. (2014), p. 652.

4 This paragraph follows Dehling, Gao, Sunyaev (2014), p. 11.

5 See Pollach (2007), p. 104.

apps of similar functionality, the high volume of apps available in the app stores[6] makes it laborious to review all of them by hand. One way to assess information privacy risks of the large amount of mHealth apps is to automate the review process of each individual app. The assessment automation can be done by downloading and analyzing the source code of each app and by tracing data leaks. Static code analysis is used in the field of informatics to analyze application source code and detect faults or vulnerabilities.[7] It is yet unclear how and to what degree the concepts of static code analysis and information privacy risk assessment can be combined in order to automate app assessment. A static code analysis could, in theory, be used to automatically assess some of the information privacy risks that mHealth apps pose. Previous research has not shown how and to what degree the combination of static code analysis and information privacy risks assessment is feasible in the field of mHealth app information privacy risk assessment and therefore the aim of this study is to explore the possibilities of static code analysis for information privacy risk assessment. This leads to the research question: How and to what degree can the information privacy risks of mHealth apps be automatically assessed? The 'degree' refers to the amount and the level of detail that information privacy risk factors can be automatically assessed.

The automated process furthermore can help to drastically reduce the effort of reviewing each individual app and can enhance the information experience users receive while looking for mHealth apps. Additionally, it exposes new possibilities for research in the information privacy risks area. The research could be conducted on providing solutions and best practices for further enhancing the information privacy risks communication of apps.

1.2 Objectives

The main objective of this study is to ascertain how and to what degree the assessment of information privacy risk factors for mHealth apps can be automated. In order to reach this objective, the following sub-objectives have to be met.

The first sub-objective is to extract information privacy risk factors from the infor-

---

6 See Enck et al. (2011), p. 1.

7 See Baca, Carlsson, Lundberg (2008), p. 79.

mation privacy practices that Dehling, Sunyaev (2016) identified and that are relevant for automated information privacy risk assessment. As a second sub-objective we will develop strategies to identify the information privacy risk factors within the source code of mHealth apps via static code analysis. This is necessary since it is yet unclear how and to what degree the static code analysis can help to identify information privacy risk factors of mHealth apps. Finally we will evaluate how well the automated information privacy risk assessment tool can identify information privacy risk factors in comparison to two human reviewers. In order to fully ascertain the degree static code analysis can identify information privacy risk factors, a manual review of the results of the static code analysis is necessary.

1.3 Structure

2. Combining Source Code Analysis with Information Privacy Risk Assessment

mHealth apps have been examined in various research studies that aim at providing insights for developers as well as users into how private information is processed. Privacy issues are the most impactful user complaint while using mobile apps.[8] This encourages research to address information privacy risks.

Research focus has been put on the technical side of information privacy breach. It has been analyzed, to what degree the data storage in internal Android log files or on the memory card within a phone or tablet poses a threat to users information privacy.[9] Technical evaluation of mobile apps even goes further into the topics of decompilation to analyze device identification or geolocation data leaks.[10] Decompilation reveals to be a feasible assessment technique for information privacy risks and data leaks.

In informatics and software development contexts, static code analysis has been used to analyze source code and provide feedback on coding styles to the users while programming or "to find defects in programs"[11]. Static code analysis provides a fast way to analyze source code[12], which makes it suitable to automate the assessment of large datasets. A further benefit of using static code analysis to retrieve information from software is that the software does not need to be executed during the analyzation process. This additionally supports the development of fast performing assessment tools that are suitable for application on large datasets of source code since there is no need to wait for the application runtime to execute the software.

Our study will use the benefits of static code analysis and apply them to the assessment of mHealth information privacy risks. It is unclear if static code analysis is a viable tool to analyze and identify information privacy risk factors. We will use the comprehensive privacy-risk-relevant information privacy practices that Dehling, Sunyaev (2016) identified[13] and try to implement static code analysis strategies to identify those risks au-

---

8 See Khalid et al. (2015), p. 5.

9 For the previous two sentences, see He et al. (2014), p. 645-646.

10 See Mcclurg (2012), p. 1, 5., Enck et al. (2011), p. 1. and Mitchell et al. (2013), p.6-7.

11 Bardas, Others (2010), p. 1.

12 See Bardas, Others (2010), p. 5.

13 See Dehling, Sunyaev (2016), p. 8-17.

tomatically. This will be a vital addition to current research, since there is yet no holistic approach to apply static code analysis to information privacy risks detection that takes an ample amount of information privacy risk factors into account.

2.1 Information Privacy Risk Assessment

2.2 Static Code Analysis

2.3 Relevant Information Privacy Risk Factors

For this thesis, we used the set of information privacy practices extracted from literature and ... by Dehling, Sunyaev (2016) as a source to derive information privacy risk factors from. Since not all information privacy practices that an app provider can include in his privacy policy may express an information privacy risk, we extract the information privacy practices that are relevant in terms of posing and expressing a potential information privacy risk. The full list of information privacy practices including a comment, wether

3. Implementation and Evaluation of an Automated In- formation Privacy Risk Assessment Tool

3.1 Implementation of an Automated Information Privacy Risk Assessment Tool

3.1.1 Download Phase

3.1.2 Decompilation Phase

3.1.3 Static code analysis Phase

3.2 Evaluation of an Automated Information Privacy Risk Assessment Tool

4. Feasibility of Automated Information Privacy Risk Assessment

4.1 The Automated Information Privacy Risk Assessment of Free Android mHealth Apps

4.1.1 Download Phase

4.1.2 Decompilation Phase

4.1.3 Static code analysis Phase

4.2 Evaluation of the Auto- mated Information Privacy Risk Assessment Tool

5. Discussion

5.1 Principle Findings

5.2 Contributions

5.3 Limitations

5.4 Future Research

5.5 Conclusion

References

Baca, Carlsson, Lundberg (2008)

Dejan Baca, Bengt Carlsson, Lars Lundberg: Evaluating the cost reduction of static code analysis for software security. In: Proceedings of the third ACM SIGPLAN workshop on Programming languages and analysis for security - PLAS '08. 2008, p. 79

Bardas, Others (2010)

Alexandru G Bardas, Others: Static code analysis. In: Journal of Information Systems & Operations Management. No. 2, Vol. 4, 2010, pp. 99–107

Chen et al. (2012)

Connie Chen, David Haddad, Joshua Selsky, Julia E Hoffman, Richard L Kravitz, Deborah E Estrin, Ida Sim: Making sense of mobile health data: an open architecture to improve individual- and population-level health. In: Journal of medical Internet research. No. 4, Vol. 14, 2012, e112

Dehling, Gao, Sunyaev (2014)

Tobias Dehling, Fangjian Gao, Ali Sunyaev: Assessment Instrument for Privacy Policy Content: Design and Evaluation of PPC. In: WISP 2014 Proceedings. 2014,

Dehling, Sunyaev (2016)

Tobias Dehling, Ali Sunyaev: "Designing for Privacy: A Design Theory for Transparency of Information Privacy Practices". 2016

Enck et al. (2011)

William Enck, Damien Octeau, Patrick McDaniel, Swarat Chaudhuri: A Study of Android Application Security. In: Proceedings of the 20th USENIX Conference on Security. No. August, Vol. SEC'11, 2011, pp. 21–21

He et al. (2014)

Dongjing He, Muhammad Naveed, Carl A Gunter, Klara Nahrstedt: Security Concerns in Android mHealth Apps. In: AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium. Vol. 2014, 2014, pp. 645–54

Khalid et al. (2015)

Hammad Khalid, Emad Shihab, Meiyappan Nagappan, Ahmed E. Hassan: What Do Mobile App Users Complain About? In: IEEE Software. No. 3, Vol. 32, 2015, pp. 70–77

Mcclurg (2012)

Jedidiah Mcclurg: Android Privacy Leak Detection via Dynamic Taint Analysis. In: . 2012,

Mitchell et al. (2013)

Stacy Mitchell, Scott Ridley, Christy Tharenos, Upkar Varshney, Ron Vetter, Ulku Yaylacicegi: Investigating privacy and security challenges of mhealth applications. In: 19th Americas Conference on Information Systems, AMCIS 2013 - Hyperconnected World: Anything, Anywhere, Anytime. Vol. 3, 2013, pp. 2166–2174

Pollach (2007)

Irene Pollach: What's Wrong With Online Privacy Policies? In: Communications of the ACM. No. 9, Vol. 50, 2007, pp. 103–108

Declaration of Good Scientific Conduct

Hiermit versichere ich an Eides Statt, dass ich die vorliegende Arbeit selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Schriften entnommen wurden, sind als solche kenntlich gemacht. Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer anderen Prüfung noch nicht vorgelegt worden.

Köln, den 01. September 2016

I hereby attest that I completed this work on my own and that I did not employ any tools other than those specified. All texts literally or semantically copied from other works are attributed with proper citations. This work has not been submitted in identical or similar form for any other exam, assessment, or assignment.

Cologne, September 1st, 2016

Curriculum Vitae

Persönliche Angaben

Name:	Thomas Brüggemann
Anschrift:	Hoferkamp 9, 41751 Viersen
Geburtsdatum und -ort:	31.08.1989 in Viersen
Familienstand:	verheiratet

Schulische Ausbildung

1997 - 2001	Katholische Grundschule Boisheim
2001 - 2009	Bischöfliches Albertus-Magnus-Gymnasium in Viersen, Abschluss: Abitur

Grundwehrdienst

07/2009 - 04/2010	Wehrdienstleistender, Luftwaffe - Jagdbombergeschwader 31 ”Boelke”, KvD für das Wachpersonal, Fliegerhorst Nörvenich

Studium

10/2010 - 03/2014	Universität zu Köln, Wirtschaftsinformatik, Bachelor of Science
10/2014 - 09/2016	Universität zu Köln, Information Systems, Master of Science

Beruflicher Werdegang

05/2010 - 09/2012	Thomas Trefz Consulting, Köln, Softwareentwicklung im Bereich Microsoft .NET
10/2012 - 10/2014	Beister Software GmbH, Aschaffenburg, Softwareentwicklung im Bereich Microsoft .NET
10/2014 - heute	Selbstständiger Softwareentwickler und IT-Berater