Summary of ”The Curious Case of the PDF Converter that Likes Mozart: Dissecting and Mitigating the Privacy Risk of Personal Cloud Apps“

1 Preface

The value of personal data is widely undisputed in today's society – at least it ought to be. According to a study conducted by the Federal Trade Comission (FTC) in 2014 nine key data brokers, companies that buy and resell personal data, generated $462 million in revenue in 2012 alone [1]. And while people mostly want to keep their private data private [2] there are still plenty of services we voluntarily give our data to – free of charge. Obvious examples for this are social networks in which users provide their personal data for the world to see. But these networks are not the only services that give their users the option of uploading their data to the internet. Personal cloud services such as Google Drive allow users to store their data in the cloud, so that they can access it from any of their devices. The security of these services has been the subject of much debate [3] but it is not only the providers of cloud storage services who are granted access to the user's data. Third party apps that act on top of Google Drive and enable additional functionality in turn require access to the user's data. Obviously, most third party cloud apps need access to some data to perform their function. For example, a pdf merging app would always need read access to your pdf files to work. Often, it is hard for users to discern between privileges an app needs and privileges that are unnecessary. While cloud service providers may very well have sufficient security measures in place, there is a potential for third party apps to circumvent these measures via unnecessary privileges.

2 Summary

The conference proceeding ”The Curious Case of the PDF Converter that Likes Mozart: Dissecting and Mitigating the Privacy Risk of Personal Cloud Apps“ by Hamza Harkous, Rameez Rahman, Bojan Karlas and Karl Aberer [**?**] examines third party apps for Google Drive that are ”over-privileged“. An over-privileged app requests more privileges (and therefore user data) than it needs to fulfil its purpose. This data can be sold to advertising providers or otherwise exploited. In order to raise awareness on the user's side, the authors suggest and evaluate new permission models. Furthermore, they introduce their own app store dedicated to user privacy, analyse the behaviour of app developers and suggest best practices for cloud storage providers (short CSP). To examine the problem of over-privileged apps Harkous et al. consider and analyse third party apps for Google Drive because it had about 240 million active users in 2014 and around 420 apps marked as ”Works with Google Drive“. But the presented results should be applicable for any other comparable platform.

Each app from any store or website that works with Google Drive specifies a set of permissions from the Google Drive API or other Google APIs and can in this way access the user data stored at CSPs. The user has to accept these permissions if he wants to use the app but can revoke the permissions later (but then of course he can no longer use the app). In order to show the distribution of over-privileged third party apps the authors examined 100 randomly chosen apps from the start page in May 2015 and reviewed them by hand. In the review process each app is linked with Google Drive and the requested and required permissions are recorded. The results are as follows: 64 percent of the apps are over-privileged and 76 percent of apps request full access to the user files (as all of the over-privileged do). This means that 84 % of the apps that want full access are over-privileged. The authors point out that it was the easiest to manually revise the apps because automation would add a lot more complexity and would be outside the framework of this work. But they remark that they are currently working on such a solution.

2.1 Suggestion of new permission models

As the next step in this paper the authors describe and evaluate three alternative permission models to the existing model of Google Drive: Delta Permissions (DP), Immediate Insights (IM) and Far-reaching Insights (FR). The DP model is based on the assumption that users are deterred from privileging an app if they are told which permissions the app actually needs and which it requests unnecessarily. For the second model, IM, the main idea is that users are deterred when they are confronted with information that can be gained from the user data by means of the unneeded permissions. Possible displayed insights could appear as an image, the location it was taken, the beginning of a text file or the name and profile picture of a colleague. The last and most complex model is FR. The theory in this case is that users are deterred if they are shown which far reaching insights can be gained from their data via the unneeded permissions. Harkous et al. present six types of insights in this category: Entities, Concepts and Topics (Names of people connected to the user and user interests), Sentiments (topics the users has positive and negative feelings about), Top Collaborators (people most interacted with), Shared Interests (interests shared with a group of people), Faces with Context (pictures of the people that occurred the most) and Faces on Map (the locations photos of people were taken).

After introducing these models, the authors present a survey to see how users react to these permission models. They looked for Google Drive users at their university with a minimum of ten text files or 20 images stored as survey participants. They found 210 persons for their study who were split into four groups: three groups corresponding to the permission models (DP 50, IM 54, FR 51 participants) and a control group (55 participants) with the original Google Drive permission interface. Participants were confronted with multiple apps, their corresponding permissions and the consequent "problems" for the permission model as described above. The participants then had to decide whether they would privilege that app or not. The results were assessed with the method of Acceptance Likelihood (the acceptance relative to the overall feedback $\rightarrow$ the less the better). The most effective method to deter people from using an over-privileged app was found to be showing them Faces with Context (~0.08) or Shared Interests (~0.09), both from the Far-reaching Insights model. The least effective were found to be Delta Permissions (~0.42) or the standard permission model (~0.39). However, the authors point out that these results may be distorted due to the choice of participants and artificial presentation of the apps. They concluded that there is a need for further studies in this field.

2.2 PrivySeal

Once the authors of the paper evaluated how an effective permission model could be designed, they decided to develop and launch their own app store for Google Drive apps. It is called PrivySeal and focuses on the protection of the user's privacy. For this purpose, there are three core features. First, apps are enriched with the information as to which of its permission requests are unneeded and the possibly gained far-reaching insights from that. Second, the user is given the possibility to search for apps according to their privilege requirements. Third, a tool is provided for developers to see which permissions they requested unjustifiably and give them a list of permissions they could have requested instead. For the year 2016, the authors claimed that PrivySeal had 1440 registered users and offered 100 apps.

2.3 Current misbehaviour of developers

After finding out that many of the third party apps request more permissions than they actually need and presenting their own app store, Harkous et al. further analysed the current practices of app developers. Therefore they investigated previously installed apps at Google Drive from their 1440 app store users (662 apps in total). For each of those, they recorded whether the app requested partial or full access to the users Google Drive data at the time of authorization (if the app changed the permissions later it was recorded as if it wanted both access types) and from which source the app originated (Google Chrome Web Store (~159 Apps), other Google Web Stores (~66 Apps), outside of Google Web Stores (~437 Apps)). The results show that developers in the Chrome Web Store tend to request full access less often (14% full vs. 48% partial vs. 40% both), while developers outside of Google Web Stores do it the most (64% full vs. 35% partial vs. 0% both). They also show that if the permissions are changed during the existence of an app, developers most of the time request more access: 94 percent of the examined apps that changed their permissions at the Chrome Web Store changed from partial to full access. This could indicate that these apps request more data without increasing their functionality. But another possible problem is the access to the users metadata. Harkous et al. performed an experiment with 200 registered users (for their app) with each more than ten text files with related topics to show the risk of metadata. The results for "General Topics" show that about 78 percent of the metadata labels match the top five interests of the user. In case of "Specific Topics" two-thirds appear in the top ten and in case of "Concepts" only 31 percent appear in the extracted data. But the other 69 percent mostly are similar and could also

be relevant to the user.

In conclusion it has to be said that enhancing the privacy in the Chrome Web Store is not enough to reach and deter developers from misusing app permissions because most apps are not in the store. Therefore alternative solutions to analyse the permissions of apps from various sources have to be found.

2.4 Best practices

The last part of the paper suggests best practices for CSPs, which should help the user to not expose his data unnecessarily. The first suggested practice is to use more granulated permissions restricted for example for specific types of files. The authors point out however, that this could possibly result in more complicated interfaces. The second proposal is to add an overview to cloud platforms where the user can see which files were downloaded by third party apps and when this happened. The third recommendation is that cloud platforms should inform users about insights that can be gained from files that are currently downloaded. The last offered advice is to add an additional API that works as a top layer for other APIs and takes care of privacy matters. This could be combined with the previously suggested practices.

3 The paper's influence

The paper "The Curious Case of the PDF Converter that Likes Mozart: Dissecting and Mitigating the Privacy Risk of Personal Cloud Apps" has reached 7 citations so far [**?**]. Two of those citations are from the paper's authors. According to Google Scholar, Hamza Harkous has been cited 97 times in total [**?**]. This particular paper is the fifth most cited one which he published. Regarding the fact that it has been published over a year ago, this low amount implies low scientific influence.

While this paper's scope is rather large and not all too specific, Harkous et al. conduct more detailed research into some of the topics discussed here in their following papers. One of these papers discusses the privacy loss caused by a user's collaborators [**?**]. Another goes into further detail on how to better present users with the data they are giving third party apps access using Data-Driven Privacy Indicators [**?**].

The authors' PrivySeal App Store does still exist but does not seem to have seen significant updates since May 2015 indicating the experimental nature of the project [**?**]. While the App Store's ambitions are interesting it is a shame the authors didn't follow through with the project.

The paper and its subject are not very controversial. Even though the results from Harkous et al. build a scientific statement, they do not challenge a deeper discourse. Neither do they build up to any revolutionary idea, which could impact the scientific community in a strong manner. This shows especially in the way, the paper is received and therefore cited from other authors. This paper is used as an example for privacy risks by over-privileged applications. Other authors now use this to clarify that there are indeed such risks which they try to tackle in their own work.

It would be wrong to claim that this paper has no value or that it would be redundant in any way. Third party cloud apps are currently no hot topic (yet) in the scientific community.

4 Reactions to the given Feedback

At first, we would like to thank all three groups that reviewed our paper. We received a good mix of feedback on both the content, form and language of our summary. Two out of three reviews had very reasonable suggestions for us although some points might have needed a little extra explanation. One of these two reviews shows our peer reviewers really took their time and read the original paper to see if we missed any important details. On the other hand, the final review we received was a little lacking. While the reviewers made it clear they were not happy with our summary they mostly failed to voice their criticism in a constructive manner that would have allowed us to improve on some of the flaws they described in our work. In the end we were happy with the feedback we received and reworked a good portion of our summary accordingly.

References

[1] Federal Trade Comission, *A Call for Transparency and Accountability*, https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf, called 2017-11-25

[2] Alexis C. Madrigal, *How Much Is Your Data Worth? Mmmm, Somewhere Between Half a Cent and $1200*, https://www.theatlantic.com/technology/archive/2012/03/how-much-is-your-data-worth-mmm-somewhere-between-half-a-cent-and-1-200/254730/, called 2017-11-25

[3] K. Ren, C. Wang and Q. Wang, *"Security Challenges for the Public Cloud," in IEEE Internet Computing, vol. 16, no. 1, pp. 69-73,*

Jan.-Feb. 2012, http://ieeexplore.ieee.org/document/6123700/, called 2018-01-06

[4] H. Harkous, R. Rahman, B. Karlas, K. Aberer, *The Curious Case of the PDF Converter that Likes Mozart: Dissecting and Mitigating the Privacy Risk of Personal Cloud Apps*, https://petsymposium.org/2016/files/papers/The_Curious_Case_of_the_PDF_Converter_that_Likes_Mozart__Dissecting_and_Mitigating_the_Privacy_Risk_of_Personal_Cloud_Apps.pdf, called 2017-12-29

[5] Goole Scholar, *The Curious Case of the PDF Converter that Likes Mozart: Dissecting and Mitigating the Privacy...*, https://scholar.google.de/scholar?cites=17319405987426194110&as_sdt=2005&sciodt=0,5&hl=de, called 2018-01-14

[6] Google Scholar, *Hamza Harkous*, https://scholar.google.de/citations?user=EzQ9nw0AAAAJ&hl=de&oi=sra, called 2018-01-14

[7] H. Harkous, K. Aberer, *"If You Can't Beat them, Join them": A Usability Approach to Interdependent Privacy in Cloud Apps*, https://infoscience.epfl.ch/record/224430/files/harkous_codaspy_2017.pdf, called 2017-11-25

[8] H. Harkous, R. Rahman, K. Aberer, *Data-Driven Privacy Indicators*, https://www.usenix.org/system/files/conference/soups2016/wpi16_paper-harkous.pdf, called 2018-01-10

[9] H. Harkous, B. Karlas, R. Rahman, K. Aberer, *PrivySeal Website*, https://privyseal.epfl.ch/#, called 2018-01-10