Deduplication as security issue in cloud services, and its representation in Terms of Service Agreements

Cecilia Wirfelt Louise Wallin
Email: {cecwi155, louwa538}@student.liu.se
Supervisor: Jan-Åke Larsson, {jan-ake.larsson@liu.se}
Project Report for Information Security Course TDDD17
Linköpings Universitet, Sweden

Abstract

The report addresses security risks that occur by the usage of online storage services, in the cloud, and how Terms of Service Agreements handle them. Especially risks that are introduced due to deduplication are considered and supplemented with the general content about user security risks in service agreements.

The report contains an introduction to what deduplication is and why it is used followed by a discussion about what risks it might introduce. There is also a section about what services that use deduplication which turns out to be two out of the three chosen ones. Next is the part about Terms of Service where each service is described alone at first and then general conclusions are presented. As presented at the end of the report, the conclusions are that security risks of this kind are seldom mentioned by providers, different services are regulated by different laws and the user should preferably have good knowledge about security risks to be sure to use the services in a proper way.

1. Introduction

Cloud Storage Services are more and more frequently used, but how many users actually understand the Terms of Service and the risks?

Probably very few: And the number of users who truly understands the concept of deduplication, how often it is used and the risks that comes with it is probably even smaller. In this report the authors will present their research on how some service providers that use deduplication treat risks in their Terms of Service Agreement. The authors will also try to detect if deduplication is performed in some of the most commonly used cloud services and discuss the impact and different risks have on different types of users. The services studied in this report are Dropbox, Drive and SharePoint.

1.1 Scope

The authors have chosen to study the services Dropbox, SharePoint and Google Drive because these are commonly used by people all over the world and are therefore interesting to see how/if they use deduplication, how they handle risks that comes with deduplication and how these aspects are addressed in their Terms of Service Agreements. SharePoint is used by Linköping University and therefore interesting for the authors.

Since this is a research about information security aspects in the use of cloud storage services the authors won't go in depth with problematic and explanations of hardware and software systems, instead the report will be limited to the problematic with security risks that users can encounter when using online storage services and how these risks are treated in the Terms of Service Agreements.

The Terms of Service Agreements or user agreements are very extensive and in this report the authors have chosen to limit the research to how the Terms of Service Agreements handle risks that can occur due to deduplication.

2. Background

In this section the authors will describe the technical concepts that later on will be discussed in the report. In this section the authors will also give example of possible risks that can occur because of deduplication.

2.1 Technical background

2.1.1 Cloud Storage Service

Today the produced data volume increases rapidly, this increases the demand for storage. Online storages are services ranging from simple backup to comprehensive cloud storage infrastructures, the later version refers to a scalable storage space made available to one or more users through the internet [1]. This services can be both charged

for in different ways or free to use [1] however the pricing is not included in the scope of this article.

2.1.2 Explanation of Deduplication

Bandwidth and storage are the two most limited resources in online storage, as described by D. Harnik et al. [1]. These two can both be saved by the concept deduplication, which is a process to prevent copies of identical files to be saved. D. Harnik et al. explains the concept by "*The term data deduplication refers to techniques that store only a single copy of redundant data*" [1].

The report is about two kinds of deduplication, source-based and cross user- deduplication. The first one because it saves bandwidth when, instead of sending a file for saving over the internet to a server, which is consuming bandwidth, a hash function is computed on the file on client side creating a so called checksum that only consists of typically 256 or 512 bits. This checksum is sent to the online storage and compared with all the previously stored material and if a match is found it is known that an identical file is already available, and in that case the file won't be uploaded again. Since the file, that is far bigger than the checksum, will never be transported over the network if unnecessarily the positive effects will be gained. The nature of a hash function is that every input gives the exactly the same output every time, and therefore it is guaranteed that a checksum available online is indeed a unique file. [1]

Cross-user deduplication is deduplication based on the material at multiple user accounts. In other words, a file uploaded to a cloud based storage service will not be duplicated even though two different users, with no access to each other's accounts, upload it. [1]

From now on the authors will refer to the combination of source-based and cross-user deduplication by the term deduplication. To sum up, deduplication saves valuable resources both for the users and for the supplier, however it introduces a number of security risks that will be described further later in the report.

2.2 What services perform deduplication

In this section the authors will discuss whether or not the chosen service providers use deduplication, how to detect deduplication and compare how deduplication is used by the different providers.

2.2.1 Method

According to Harnik et al. [1] services that use deduplication are not hiding it but express it in the file history log or even express it in the direct communication like status messages to the users but, as the article mentioned, this is of course based on the interface. Also the time consumed in an upload event compared with the bandwidth by the client revealed deduplication. Naturally, based on the description of pros with source-based deduplication, a method that will work for detection in any interface is to monitor the network traffic to register if the file is transmitted. [1]

When Harnik et al. tested the above mentioned detection methods, they were surprised to see that no one of the services tried to hide the fact that deduplication had occurred. Especially the fact that it was outspoken in status messages or file history was remarkable. [1] The fact is that if it was made impossible to detect the occurrence of deduplication, the integrity risks introduced would be eliminated.

These above mentioned methods will be tested (to limited extent) on some services that supposedly use deduplication and whether or not the tests succeeded and if the results were as expected will be presented in this subsection.

2.3 Risk due to deduplication

As mentioned before, security risks can occur because of deduplication. For example since the data is only uploaded once, there is a potential risk of data loss. When a storage service use cross-user deduplication users could also be concerned with the integrity of data since they might not actually save the data but get a pointer to someone else's copy and if there should be a hash collision that other copy might not be what is expected. To explain more closely how deduplication could affect users in practically three examples of attacks will be described.

The first example is that if someone has a specific sensitive file and suspects that another user also has that specific file. Then the first user can check if it is true by looking for deduplication. This can, for example, be used by the police. Say that they upload a specific sensitive file with illegal material to a cloud storage service and find that deduplication occur, than they could go to the service and get information about the people who have access to that file. [1] This could of course be a very good thing, but it could also lead to someone being falsely accused if an innocent is set up by a criminal or, depending of the Terms of Service Agreements, malicious users might be able to get information about other users.

Another example is that an attacker could use a *brute force* attack to find sensitive information. This can be done

if an attacker has part of a specific sensitive file, for example a document template with a PIN code, by uploading a number of files with different combinations of the PIN the attacker can find the right one, if the document is completely identical to the original apart from the PIN. Deduplication will only be detected when the document with the correct PIN is uploaded. [1]

The third example is if a service use cross-user deduplication an attacker could exploit it to create a covert channel between some malicious software and a remote control centre run by the attacker. [1]

2.3.1 Evaluation of risks

Some services might have a Terms of Service Agreement that states that no information about users will be relieved to no one and then the confidentiality of the data won't be affected, but criminals might get away their illegal activity. How services handle confidentiality is one of the things that will be further investigated in this report.

Considering the second example attack above it might not be very likely that a private user upload such sensitive information to an cloud storage service, and if they did an attacker would have to know a lot about the document to be successful with the attack. However if deduplication is used within a company and a malicious user gets access to the system it might be a risk of more importance, but that won't be discussed in this report since it discuss cloud storage services and the information risks when used by private users.

3. Deduplication in Terms of Service Agreements

In this section the authors will discuss how the chosen service providers address the previously mentioned risks in their Terms of Service Agreements and the question "*How does Terms of Service Agreements handle the deduplication problematic?*" will be answered. The authors will also bring up the method used to find out if and how the risks are addressed in the agreements, as well as how it can affect the users.

3.1 Does the services perform deduplication? - Test results

In this subsection the results from the previous tests will be discussed and the different service providers will be compared.

3.1.1 Drive

Research shows that Google Drive doesn't use deduplication. An interesting thought regarding this is what the reason for not using deduplication could be, since deduplication, as mentioned above, has some advantages even though it also enables new attacks. One reason for not using could be that because the service allows live operating environment

3.1.2 SharePoint

According to a help centre blog that Microsoft controls, SharePoint uses Shredded Storage which conceptually can be called deduplication for documents within a folder. [7] The architecture, Shredded Storage, on which SharePoint is developed, makes sure that only changes to a file will be stored. This is called "*an effect similar to deduplication or single instancing*". [8] The authors will consider this single user-deduplications since a folder can be shared among a number of users and then a file that already exists won't be uploaded again but only when it is obvious to the user. Also the fact that a copyright owner can upload a file, check weather deduplication occurs to discover if a copy of an identical file exist in the storage will not be possible outside a shared folder that the copyright owner is a member of this shared folder which is an additional argument to why this is considered single user deduplication. Since SharePoint is an internet based service it is reasonable to assume that the check that confirms if the storage contains a file identical to the one that is about to be stored, is performed online. That would imply that no source-based deduplication occurred and therefore the advantage of saving bandwidth would not be gained.

3.1.3 Dropbox

Tests revealed that Dropbox uses deduplication for private users [1], however only source based deduplication and not cross user deduplication. Since a business license for Dropbox would be needed to test if Dropbox uses deduplication for their business services and the authors were not able to get hold of such license, due to license cost, such a test could not be performed. Since the service is known to perform deduplication for private users the authors have, for argument sake, concluded that the providers use deduplication for business services as well.

3.2 Term of Service Agreements

In this subsection the authors will present what they looked for when reading different agreements and explain how it is relevant for the mentioned risks.

3.2.1 Drive

Since the research show that Google Drive doesn't use deduplication the risks discussed above won't be an issue and therefore the providers Term of Service Agreements has not be closely studied.

3.2.2 SharePoint

The first conclusion is that the Terms of Service Agreements of SharePoint is not easy to find, not even for a registered, logged in user. This is problematic for the individual user since he or she cannot find specific information about what they have signed up for but for an organization the problem might not be as important. This is because as an organization probably have a specific agreement designed for their particular use of the service and a responsible person who manages it and keeps user terms available. The later case is how Linköping University distributes SharePoint to its employees and students uses SharePoint.

Secondly, well stated in the terms of services is that SharePoint apply the laws in the country where the licenses are acquired. [4] With that said, it is assumed that it does not have to be the country the customer or user is located in that determines what laws that needs to be followed. For example if an user acquires the license abroad, in a country that is not the location they intend to use the license, it most likely means that when the user gets back home it is still the laws of the abroad country that will be applied. That is relevant since laws regulate the interpretations of the terms of services and what a user is allowed to do to might change. This also stresses the need for a commercial user (such as a company) or organizations who treat sensitive information need to have a suiting contract before using a service that apply a law they are not aware about.

The Agreement contains a comprehensive disclaimer that states

"*To the extent permitted under your local laws, Microsoft excludes the implied warranties of […] non-infringement*". [4]

This means that Microsoft will not take responsibility for intrusions and in other words, if occurrence of deduplication shows that there is copyright protected material the SharePoint storage and intrusions to user accounts are performed, Microsoft does not have any obligations to react. How and if an attack like that could be put through towards multiple users accounts to find out whether anyone has a copy of a certain file is not the topic of this section, however the protection of private contents is and it is interpreted as if Microsoft does not consider it their concern.

How SharePoint takes part of and collects information about their users, information such as IP-addresses, operating system and so on, is expressed to be in order to develop features and future products.

3.2.3 Dropbox

Dropbox uses different agreements depending on if the user is a private user or a business user. For private users there is a "Privacy Policy", this policy states, among other things, that Dropbox takes responsibility for keeping the users data safe. It also tells the user what Dropbox will store and who will be able to access the stored data. [8] An interesting aspect is how Dropbox has chosen to formulate why they store particular data, namely that they are claiming that it is good for the user. The Privacy Policy states

"*When you use our Services, we store, process and transmit your files [...] and information related to them. [...] This will make it easy for you to do things like share your stuff, send emails, and invite others to use the Services."* [8]

The tone of this statement says that it is a good thing for the user that Dropbox store information related to the users uploaded files which the user actually has no choice in accepting weather or not Dropbox should be allowed to store. Dropbox continue to explain why this is good for the user but they do not state whether or not they take any benefit from it. Depending on how the user uses the service the kind of information stored about him to help Dropbox "improve usage of the service" might not actually be relevant to this particular user. Still the user cannot deny Dropbox to store the extra data. Such data could be information about the device being used such as what site the user visited before logging onto Dropbox and location information. Dropbox also states in the Privacy Policy that in particular occasions a third party can get access to the information stored about the user, again this is namely to improve the service (apart from a case where authorities are involved because of illegal activity). The risks that are not treated in the Privacy Policy are if they take any responsibility if the third party misuses the information in anyway.

The Privacy Policy also discusses how Dropbox relate to the law, stating that they act according to the Californian law apart from the situations where there are law principle conflicts. [6] This could be an issue for a user who is not familiar with the Californian law. Say for example for

someone in Sweden who is used to being protected by PUL, the Swedish Personal Data Act, might not realise that when creating a Dropbox account different laws are considered. Furthermore Dropbox states that they can disclose information to a third party for the following reasons

"[...] if we determine that such disclosure is reasonably necessary to (a) comply with the law; (b) protect any person from death or serious bodily injury; (c) prevent fraud or abuse of Dropbox or our users; or (d) protect Dropbox's property rights." [8]

Regarding the probability that many users who are not from California, and presumably even users who do live there, do not know exactly what acts are against the law the statement above might lead to a lot of problems for the user, without them realising that they are doing something illegal. This of course could be discussed if it really is an issue, but laws are different in different countries because we do not always agree with each other and that could mean that someone in Sweden would get into trouble for something that the Swedish law does not regard as illegal.

Another interesting discovery that has recently been made is that leakage has occurred when users has clicked on links within documents or links to shared files are accidentally put into the Google search box. [2] The article that discusses this leakage suggest to Dropbox users to upgrade to the Business version, since in that version the user can restrict access to shared links. In the agreement for business users Dropbox lay a lot of the responsibility on the organization (costumer) that have bought the license. [9] This allows the users to protect themselves from risks that they regard as important for them. For example a malicious end-user within the organization might be able to use a brute-force attack to exploit the system and Dropbox would take no responsibility of such events, but the user would be able to administrate protection within the organization. An important aspect to remember concerning this is that when the organization accepts the Term of Service Agreements they have to be aware of the risk, evaluate the impact and know how to protect themselves from the risk.

3.3 General aspects in Terms of Service Agreements

In this subsection the authors will discuss how the general results that relate to more than one service, found in the previous subsection could affect the users. A comparison between different service provider agreements will also be presented.

Generally the Term of Service Agreement does not explain what risks the users are exposed to when using the service, it only bring up what rights the user have and what the provider is obligated to take responsible for. For example the Term of Service Agreement might say that the provider is responsible for making sure that data won't be lost or get into the wrong hands. However it does not describe against what or how they protect the user. Since an attacker usually can find many different ways to hack someone's account this could make a big difference, especially considering deduplication since an attack which is made possible due to deduplication can be very hard to detect. Just because the provider states that they make sure that data is not lost or misused it does not have to mean that some malicious person is not able to get hold of the data, since it does not ensure that the provider discover every attack. The conclusion that can be drawn from this is that even though the provider promises to take responsibility for lost or misused data, the promise cannot give a hundred percent assurance to the user since it is likely that the provider might miss an attack.

When an organization or a company uses online storage through a contracted provider a customer specific contract can be agreed upon. The reason for this can both be specific needs in performance or security or depending on the provider claiming payment from profitable companies for services that are free of charge to private users. Linköping University, an authority that outsources cloud storage to a service provider. The official document LiU-2012-00700 "The use of cloud services - A Resolution" (translated) [5] from LiU confirms the need for an agreement fulfilling above mentioned criteria. This kind of contract is used in every situation where the university uses outsourcing of any service to make sure that the right needs are met. If this use of specified contracts can be extended to cover all organizations and companies, this proves that the service providers of online storage places great responsibility on the users themselves regarding performance and safety. This underlines the importance for any organization to work with security and have knowledge about both security and risks.

In all the studied cases Term of Service Agreement gives the provider rights to collect information about user habits such as operating systems used and web-pages visited before entering the service. This is used to justify an act (collection of information) that has nothing to do with the service and an ignorant user who does not read the terms might have no clue that this is being done. The question if it is ok to do this will be answered when the errand is tried in court and a resolution is made. However, today the term "By using this service, you accept our terms of agreement" has grown in popularity and gives 2,4

billion hits on google making users custom to a situation where they have limited power and might not question occurrence of events non-related to the purpose of the service. A second question is how this information is kept by the provider, something that is not mentioned at all in the agreements. This would help users determine whether this storage also would introduce risks. For example, if the information about what web-pages a user visits in connection to using the service and the information was stored in a non-protected database and a malicious user could get hold of the web history of a specific, innocent user and trace him on the different pages collecting a lot of privacy information that could be used in order to harm.

Also these online storage services are global services which introduces the problematic of what countries laws that should be applied. In the studied cases both the law from where the provider is stated (USA) and the one where the license is acquired occurs and as a user you should take this in consideration, especially companies and organizations.

4. Conclusions

The research made by the authors gives reason to believe that deduplication occurs in at least some cloud storage services. Furthermore the authors have discussed how users are protected from risks that could occur because of deduplication in the providers Term of Service Agreements. The terms that has been studied has shown that providers are quite similar and bring up more or less the same bullet points. The providers take responsibility for keeping the stored data secure but they do not address how they do it, which means that the users cannot know what kind of attacks they are protected from and what attacks they are not protected from. This kind of information could be important for the user since it is very unlikely that the providers are protecting the data from every possible attack.

Another aspect that has been discussed is that the providers follow different laws. Some apply the same law for all users no matter the location (for instance Dropbox who rely on the Californian law) and some apply the law that is enforced in the country where the license was purchased. This could mean a big difference for the users since laws in different states and countries state different things regarding illegal activity.

To summarize, users themselves have to know research what risks they could be exposed to. Risks with storing data in a cloud storage service is not something that a user can find in the providers Term of Service Agreement, and so even if the user would read the terms it probably wouldn't do much of a difference unless the user has a lot of knowledge about risks regarding this kind of storage or if the user did quite an extensive research.

References

[1] D. Harnik, B. Pinakas, A. Shulman-Peleg, "Side Channels in Cloud Services: Deduplication in Cloud Storage", Security&Privacy, 2010 (vol. 8 no. 6 pp. 40-47)

[2] http://slashdot.org/story/14/05/07/009248/Dropbox-and-box-leaked-shared-private-files-through-google, accessed 2014-05-09

[3] http://www.securstore.com/blog/risks-of-data-de-duplication/, accessed 2014-03-27.

[4] Terms of Service Agreements, section 8b. http://www.microsoft.com/en-us/download/confirmation.aspx?id=38417, downloaded 2014-05-08.

[5] "The use of cloud services - A Resolution" (translated) http://www.liu.se/om-liu/organisation/us/Protokoll_beslut_2012/2012-06-11/1.353158/09Anvndningavmolntjnster-ettprincipbeslut.pdf, downloaded 2014-05-08.

[6] https://www.Dropbox.com/terms, accessed 2014-05-06

[7] http://social.technet.microsoft.com/Forums/SharePoint/en-US/18cfac66-1ed8-4a96-814b-25319b0f1686/deduplication-in-SharePoint-2013?forum=SharePointgeneral, downloaded 2014-05-08.

[8] https://www.Dropbox.com/terms#privacy, accessed 2014-05-06

[9] https://www.Dropbox.com/terms#business_agreement, accessed 2014-05-06