NordDesign 2014
August 27 – 29, 2014
Espoo, Finland / Melbourne, Australia

GDriveMonitor: a Tool for Analysing Information Sharing Behaviour in Virtual Engineering Design Teams using Google Drive

Cornelius Illi, Franziska Häger

Hasso Plattner Institute, Potsdam, Germany
cornelius.illi@student.hpi.uni-potsdam.de, franziska.haeger@hpi.uni-potsdam.de

Abstract
Prior applications to monitor and analyse the information sharing behaviour of virtual team environments are often based on proprietary Computer-Supported Cooperative Work (CSCW) technology that does not find its way into the working routine of the modern *cloud workers*. To not limit research to artificial setups, we present *GDriveMonitor* a research tool build on top of *Google Drive* - one of the most used file storage services at the time of writing. As prior work on customisable service platforms, did not satisfy the requirements for analysis and evaluation within *Google Drive*, the necessity for a new tool was given. Possibilities and limitations are presented and discussed, as well as its application within a global engineering design course.

Keywords: virtual teams, information sharing, engineering design, computer-supported cooperative work

1 Introduction

According to Forrester Research's US Telecommuting Forecast by 2016 43% of the US workforce will, at least occasionally, be working remotely[1]. This trend towards increased virtual working environments was mainly caused by a paradigm shift in software development from self-hosted, proprietary solutions to Hardware (HaaS) and Software as a Service (SaaS), There are a lot of reasons for its rise and the advantages of cloud-based solutions [1], which explains why all major IT-companies responded with the development of new products, e.g. Microsoft with *SkyDrive* and *Office365*, *Google* with its *Google Apps Toolsuit* (GAT) , Apple with *iCloud*, Adobe with the *Adobe Cloud* or SAP with the SAP HANA Cloud Platform to name a few. These new solutions highly affect the way we work and open new possibilities for team setups across offices, companies, countries or time-zones. For collaboration in virtual environments it is therefore inevitable to find best practices. Sharing information, which can be seen as a precondition to the creation of Shared

[1] US Telecommuting Forecast, 2009 To 2016: http://goo.gl/QeOoyd

Understanding (SU) , is an essential requirement and has been subject to prior research [2] [3]. New Product Development (NPD) is also being executed in virtual setups, increasing the complexity for managing people, processes and knowledge. Knowledge that is often of "transient utility" [4]. In order to define best practices and motivate the creation of better solutions, it is necessary to gain insight into working behaviour of virtual teams. This paper will introduce a research software for monitoring and analysis of virtual teams using *Google Drive*[2] to share information and collaborate.

2 Motivation

In 2004, Stanford's ME310 Engineering Design class expanded to a course for global innovation [5], where multidisciplinary, multicultural and multi-purpose teams invent new products and services throughout a period of three quarters. ME310 has previously been subject to research on many different topics, also on tools for CSCW collaboration support [6], yet little knowledge exits on best practices for sharing knowledge and the creation of SU within engineering design teams as problems persist [7]. Throughout two surveys and several interviews with course participants, convenience was found to be one of the most influential factors for choosing a certain technology for collaboration, communication or coordination. By the end of the first quarter of this year's course, all of our student teams had switched from other file sharing service like a self-hosted *OwnCloud* or *Dropbox* to *Google Drive*. What Grudin et al. [8] already defined in 1994 by the "critical mass" problem, is something that we could confirm as technology use either being convenient or not. Most of the students already had Google accounts, e.g. for *Google Mail*. Having an account includes access to services like Youtube, Picasa, Google+, Google Hangouts and Google Drive. Most of these services have also been used before for private purposes.

However, interviews revealed that students found it difficult to handle the vast amounts of information that are constantly being shared. This motivated us to create a tool that could both benefit our students for support and us researchers for monitoring and evaluating there activity. Although, we are especially concerned with the needs for collaborative conceptual design, the presented tool is build as a general purpose monitoring tool. Furthermore, the GAT, i.e. all previously mentioned services, has not been subject to research in the field of CSCW, although it can be classified as such in respect to Johansen CSCW Matrix or even newer approaches [9]. Google provides tools for Information Sharing with Google Drive, Google Docs, Google Sites and Google Groups, communication capabilities through Google Mail and Google Hangout and team coordination via Google Calendar. Information sharing and communication services can either be used in synchronous and asynchronous ways. As the GAT is created as a platform, it offers great evaluation capabilities through its sophisticated Application Programming Interface (API), thereby allowing researchers to monitor and evaluate as well as create new solutions on top of its products.

3 Related work

There is little research about CSCW in conjunction with SaaS solutions, although the development of most commercial groupware applications already went or is aiming in this direction. Forrester Research named *cloud deployment* to be one of the 10 most important technology trends in 2013 [10]. To explain why there is little research in this field, we have to first look at the definition and differences between CSCW and groupware. Greenberg defined CSCW as "the scientific discipline that motivates and validates groupware design" [11]. By the time the statement was made, research and industry might have been progressed at similar

[2] Google Drive: https://www.google.com/drive/using-drive/

pace. As the technological foundation for SaaS solutions was not laid out, academia and industry both developed their own proprietary solutions that mostly remained black boxes to the opposing group. With the rise of cloud-technology and the consecutive development of SaaS solutions, two things changed: (a) companies are striving for platforms offering powerful APIs for extendability instead of products and (b) academia fell behind industry as more and more of the established IT-companies as well as startups around the world are creating services with the intention to connect people in new ways, either to improve collaboration, communication or coordination. This opens a new field for research possibilities as existing groupware can be used as subject of CSCW research, either by monitoring and analysing users through the use of APIs or by building services on top of these platforms. And there are numerous reasons why doing so is reasonable as these products are robust, feature-rich, are optimised for high usability and have a huge existing user-base. Furthermore, as previously stated, their usage already seems to be convenient to a high number of users, which reduces the costs for adaption. An in-depth analysis of the observed developments however is not within the scope of this paper. Instead, existing research that aims in this direction will be evaluated. *AnalyseD* [12] [13], for example, can be used for this purpose as it allows for interaction with REST-ful APIs. *AnalyseD*, and its precursor *d.store* [14], have since been used to evaluate the communication patterns of numerous collocated software engineering and virtual engineering design teams. E-mails, activity on wikis and in software-repositories (SVN, Git) have been evaluated in order to predict team performance. However, the heterogeneity of resources uploaded or edited on *Google Drive* required the development of a new solution (see 4.4 for details). Besides, there are numerous tools for CSCW with special support for collaborative conceptual design that have been developed over the past two decades as evaluated in [15]. However, they are all of proprietary nature and rely on client-based software, that is either out-dated or platform-dependent, which render them to be of little use for current collaborative working requirements.

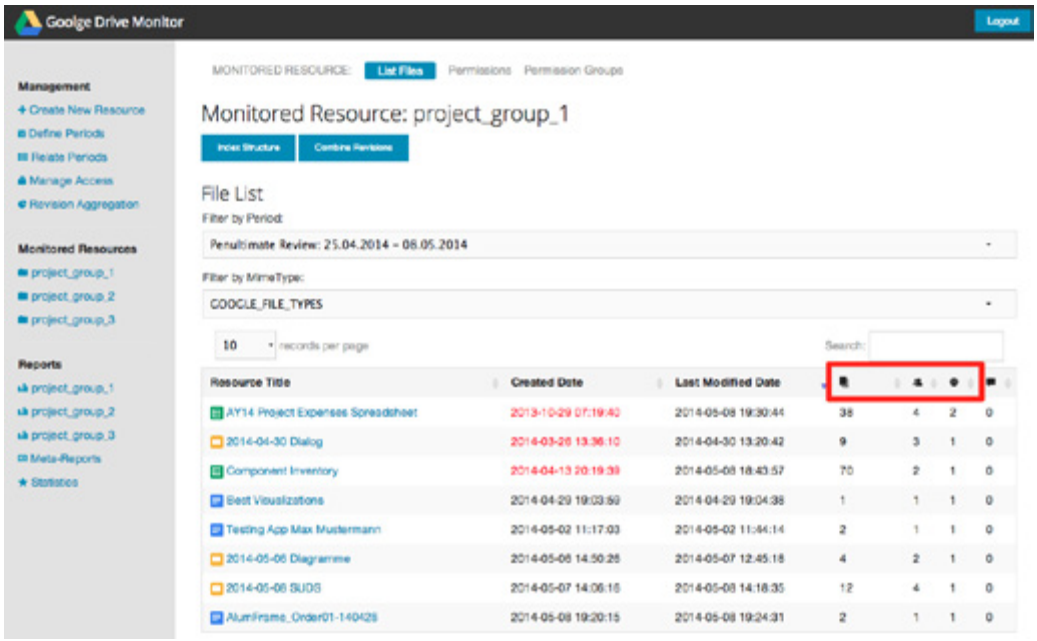

Figure 1: The GDriveMonitor Web-Interface.
Highlighted columns: number of revisions, participating users and sub-groups

4 The tool: GDriveMonitor

GDriveMonitor is a tool for monitoring and evaluating activity on *Google Drive*. It collects and stores meta-information about every resource uploaded to a folder in *Google Drive* that has been set-up for monitoring, querying the *Google Drive API*[3]. In the following sections possibilities and limitations of the API, its integration into the tool as well as possibilities and limitations for analysis of the collected data will be described.

Google handles every object in Google Drive as a file, that is connected to a parent object, either a folder or the root. Folders by itself in Google Drive are treated as files with another Multipurpose Internet Mail Extensions (MIME) type: *"application/vnd.google- apps.folder"*. In the following, the word resource is used when talking about either files or folders.

Availability The source code is published under GPLv3[4] and freely available on Github: https://github.com/cornelius-illi/gdrive-monitor

The Google Drive API Google Drive is created as a platform, allowing developers to create custom services upon its applications. Google offers a sophisticated API, that our software uses to handle authentication, authorisation and the collection of metadata about resources.

4.1 The Crawling Pipeline

After you authenticated and authorised *GDriveMonitor*, you can set up a new folder for monitoring. Once set-up the software will asynchronously index the whole structure. This process is repeated automatically in periodic intervals in order to keep track of the on-going changes. The following tasks are performed each time:

1. recursively get all resources inside the monitored folder and update their metadata
2. for each resource
   (a) fetch and store revision list
   (b) fetch and store comments
   (c) create activities by aggregating revisions that have high time density
3. if a revision has been created by a user unknown to the system, create a new permission record

Detailed information on the available data will be presented in the following sections.

4.2 Requirements

Every other week the global engineering design teams within ME310 get different assignments, e.g. for benchmarking, prototyping or the creation of presentations and the documentation. Therefore it was required to define different periods and group them according to the type of work that is being done. Furthermore, Google stores the ID of the last modifying user for each file and revision. As the global teams are divided in local subteams,

[3] Google Drive API: https://developers.google.com/drive/v2/reference/
[4] GPLv3 License: http://www.gnu.org/licenses/gpl-3.0.txt

each operating in a different country, it was also necessary to group permissions. This allows to query information about different working styles by each sub-team.

4.3 Files

For each file, the title, MIME type, creation date and the date of the last modification is stored. Thereby it is possible to track over which duration a team worked on a particular file. In our test-setting some teams worked on files over the whole period of time, while others created a new file for each version, as they did not know that *Google Drive* is keeping snapshots of every version of a file, no matter if it is uploaded to Drive or updated, e.g. in Google Docs.

4.4 Revisions

Revisions are snapshots of files at a certain point in time. Each file has at least one revision. The latest revision is called the head revision. No matter when a resource has been shared with a user, the complete previous revision history will be available to him/her. Storing revisions is useful, as each revision stores the Identifier (ID) of the editor and hence it can be monitored how many people work on a file or how many files have been edited in collaboration. One main point to understand, when talking about revisions, is that they mean different things for different types of files. When looking at *Google Docs* file-types (documents, presentations, spreadsheets, drawings, forms), a revision is an aggregation of consecutive changes made by a single-user. Google automatically saves changes and thereby frequently creates new revisions. How and when a new revision is created is not specified. Hence, revisions record the beginning and the end of a working session for *Google Docs* file-types. For documents uploaded to Drive, the revision usually marks the end of a session, depending on the users behaviour. Problems for analysis are discussed in detail in 4.4.1. As Google automatically purges revisions in order to optimise disk usage[5] and automatically deletes old revisions after 30 days or 100 revisions[6] periodical crawling is advised to keep track of all changes.

4.4.1 A homogenous view on revisions through the creation of Activities

During the analysis of the collected data we determined that revisions do not provide a homogenous view on a user's activities. The used application, e.g. *Google Docs* or a native application, as well as the individual working styles result in different patterns, that hold the potential for misinterpretation when analysing the data. Furthermore, we discovered that often batch-uploads are done, e.g. when uploading pictures, icon-sets or HTML-based documentations. Therefore we created a classification of discovered distinct activities that provides a homogenous view on the collected data-points. The attributes of each class can is shown in table 1. The time constraints relate to the time distance between two revisions using a sliding window.

[5] https://developers.google.com/drive/web/manage-revisions
[6] https://support.google.com/drive/answer/2409045?hl=en

Table 1 Classification of revision-based activities

	Action	Batch Upload	Working session
Revisions	1	n	n
Resources	1	n	1
Permissions[7]	1	1	1
Time Constraints	-	10 sec	16 minutes

For *Working Sessions* the optimal threshold has been determined by calculating all possible *Working Sessions* for thresholds between 3-40 minutes. Then the number of revisions in between the current and the following threshold value was counted and plotted. Based on the data an ideal threshold of 16 minutes has been identified. We compared this number to the duration of real-life events like restroom-, smoking- or coffee breaks that usually take between 5 - 15 minutes and found the number to be appropriate. Figure 2 shows the result of such an aggregation. The chosen file had 152 revisions, that have been aggregated to 9 Working Sessions.

4.5 Detecting Collaboration

Collaboration patterns can be determined by extending the definition of a *Working Session* (see table 1). A *Working Session* that has revisions from n > 2 permissions, i.e. different users, can be classified as a Collaborative *Working Session*. As it is possible to group permissions, Global Collaborative *Working Sessions* can be detected, when permissions from different permission-groups are present. Furthermore, Google allows to place comments in its Google Docs files. Comments are thread-based, have a context (marked passage) and a status that can be either "open" or "resolved". It is even possible to put comments to PDF-files when opened with the *Google Drive Viewer*.

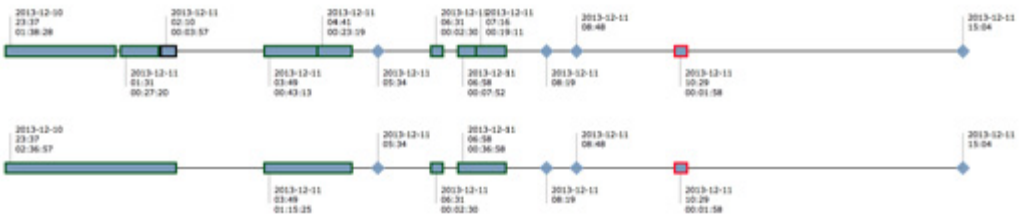

Figure 2: revisions aggregated to working sessions: 8 and 16 minutes threshold.

4.6 Limitations

Google offers the possibility to download each revision, allowing to compare changes. However, comparison only works well for text-based formats. Although all *Google Docs* file-types can be downloaded as plain-text a comparison is difficult, as a lot of information is lost, e.g. when images are inserted into a presentation or the formatting of a passage is changed. These changes will not be visible. Furthermore, Google has restrictive privacy settings, that

[7] A permission on Google Drive is a reference to a user.

limit access to its user-data. The email address of a user is only shown when his/ her Google+ profile privacy settings allow exposing it. For comments only an author-name (not the *permissionID*) will be presented, which limits the possibility to map comments to permissions. For ME310 it was common, that a single person had access to a folder through multiple accounts using different email addresses. The display name can be equal, however then the *permissionID* is different. Therefore, at the moment there is no mapping of comments to permission-records. Besides, there is only limited possibility to track if a document has been viewed by a user. Although file-resources have an attribute *lastViewedByMeDate* it only works for *Google Docs* files. Other file-types could have been downloaded by the Google Drive Desktop Application and viewed using your system applications, which is probably the reason why Google does not track views on them at all. Retrieving all views on *Google Docs* resources is also less practicable as one can only view his/her own views. Tracking the whole team requires the whole team to setup *GDriveMonitor*, which poses to be a barrier.

4.7 Reports

GDriveMonitor includes a reporting module that is creating reports grouped by the defined period-groups. For ME310 we defined four groups: prototyping, benchmarking, documentation and presentation-preparation. Report-Metrics can be of different types: general, permission-based, and permission-group based. Furthermore, a module for comparing metrics between projects as well as a system wide statistics module are available. An easy programming interface is provided allowing to extend the tool with own metrics.

4.8 Future Work

All possible ways to collect meta-data about activity on *Google Drive* through the use of the API have been investigated. The future work will be mostly concerned with drawing conclusions from the patterns we detect, which requires the development of powerful metrics. Therefore it might be useful to create a more holistic view on information sharing, e.g. by monitoring other services that are used for communication. An integration with *AnalyseD* might be of advantage.

Furthermore, other applications for the tool are being investigated. For our student teams services that increase transparency and lowers information overflow will be developed in the future.

5 Case-Study on information sharing at ME310

The sharing activity on Google Drive of three student teams, consisting of a total of 22 students spread across four nations have been monitored over a duration of six month. As can be seen in table 2 most of the files just have one single revision and most of them are images. The number of files does not split equally between the three teams: one team has around 10% of the files, where the two other teams nearly equally split the rest.

Table 2 Statistical data on files and revisions

Metric	Count	Percentage
Number of files	5271	
Number of revisions	9674	
Number of Google Docs	431	8.18%
Number of revisions from Google Docs	3818	39.47%
Number of images	3871	73.44%
Number of images with single revision	3805	72.19%
Number of files with single revision	4830	91.63%

There are some conclusions that we could draw from this data:

F1: Static Nature of Files Most files are of static nature. They are uploaded once and are never changed. Their number of revisions is 1.

F2: Google Drive is multiple purpose It is both used as a silo for information as well as a tool for collaborative work.

F3: Revisions are biased Although only 8,18% of the files are Google Docs, they account for 39,47% of the revisions, which strengthens the point that revisions mean different things for different file types. As Google Docs are edited online and auto-saved every couple of seconds, resulting in new revisions, revisions document to the working process on a file. Other documents, e.g. proprietary office files, are saved only when the user saves manually. This often happens at the end of a working session or several times in between, depending on the users individual behaviour.

F4: Detection of collaboration and parallel working activity: As the user ID of the last modifying user is stored within every revision it is possible to see, what file has been worked on globally or even if there have been parallel working sessions across national borders (e.g. as can be seen in figure 1 on page 4).

6 Discussion

The presented research tool *GDriveMonitor* is capable of real-time monitoring and analysis of team activity on *Google Drive*. For teams that exclusively make use of *Google Drive* as file-storage and for collaborative working, it provides useful insights on what and how information is shared and the individual patterns of people, local and global teams. However, it does not provide a holistic view on sharing information. Information is shared through a variety of channels in synchronous and asynchronous ways. Our student teams e.g. use mailing for communication with the liaison and the teaching teams, Facebook groups and chats for internal communication, direct conversation, phone calls and *Google Hangouts* for video-conferencing. Information sharing happens a lot through communication and often not all decisions and discoveries are persisted within a document at the time they are made or at all. For our research within ME310 we periodically conduct surveys and informal interviews in order to correlate findings within *GDriveMonitor* with statements of our students. One discovery that we made, was that one team switched back to using *Microsoft Word* for the

creation of their second documentation, as someone accidentally deleted contents while creating the first documentation, which could only partially be restored. The fear that this could happen again changed the working behavior of the whole team. Before the documentation was done in parallel with the use of comments to give feedback. Afterwards everyone created its own file that was composed to one document by the overall responsible. Every version of the file was then stored with a different name, adding a version at the end of the file, as people were insecure, wether previous revisions could actually be restored when needed. Being able to see and analyse the processes of how collaborative work is done and correlating them with survey data and evaluation of the results, allows us to better understand what works and what does not. This way we hope to derive best practises for global design teams in the future.

7 Conclusion

In this paper the research tool *GDriveMonitor* has been presented, that allows to collect meta-date from Google Drive and thereby provides means to monitor and analyse individual and team activity on *Google Drive*. In contrast to existing, more customizable solutions *GDriveMonitor* is build-upon the specific requirements for monitoring activity on *Google Drive*, that derive from the heterogeneity of the stored resources. A solution to overcome this problem has been presented by the introduction of working-sessions from aggregations of revisions. Furthermore, the possibilities and limitations of building a monitoring tool on top of *Google Drive* have been discussed. Future work, will investigate patterns and metrics for evaluation of collaboration behaviour as well as prototypes that support team activity.

References

[1] M Miller. *Cloud Computing: Web-Based Applications That Change the Way You Work and Collaborate Online*. Que, 2008.

[2] P Hinds and S Weisband. Knowledge sharing and shared understanding in virtual teams. In *Virtual teams that work: Creating conditions for virtual team effectiveness*, pages 21-36. 2003.

[3] A Malhotra, A Majchrzak, Robert Carman, and Vern Lott. RADICAL INNOVATION WITHOUT COLLOCATION: A Case Study at Boeing-Rocketdyne. *MIS quarterly,* 25(2):229-249, 2001.

[4] A Majchrzak, RE Rice, and N King. Computer-mediated inter-organizational knowledge-sharing: Insights from a virtual team innovating using a collaborative tool. *Information Resources Management Journal*, (2):44-53, 2000.

[5] T Carleton and L Leifer. Stanford ' s ME310 Course as an Evolution of Engineering Design. (March):30-31, 2009.

[6] W Ju, A Ionescu, L Neeley, and T Winograd. Where the wild things work: capturing shared physical design workspaces. Proceedings of the 2004 *ACM conference on Computer supported cooperative work CSCW* 04, 2004.

[7] K Kelly, E Connolly, M Culleton, P Weldon, and R Barrett. Issues of collaboration within global project teams.

[8] J Grudin. Groupware and social dynamics: Eight challenges for developers. *Communications of the ACM*, 1994.

[9] V.M.R. Penichet, I. Marin, J.a. Gallud, M.D. Lozano, and R. Tesoriero. A Classification Method for CSCW Systems. *Electronic Notes in Theoretical Computer Science*, 168:237-247, February 2007.

[10] J Hoppermann, P Hamerman, and G Lawrie. The 10 Most Important Technolog Trends In Business Application Architecture Today, 2013.
[11] S Greenberg. Computer-supported cooperative work and groupware: an introduction to the special issues. *International Journal of Man-Machine Studies*, pages 133-141, 1991.
[12] T Kowark, M Uflacker, and A Zeier. Towards a shared platform for virtual collaboration analysis. *Design Thinking Research* (2012), 2011.
[13] T Kowark and H Plattner. AnalyzeD: a shared tool for analyzing virtual team collaboration in classroom software engineering projects. *The 2012 International Conference on Frontiers in Education: Computer Science and Computer Engineering*, 2012.
[14] M Uflacker and P Skogstad. Analysis of virtual design collaboration with team communication networks. *Proceedings of ICED 09, the 17th International Conference on Engineering Design*, Vol. 8, (August), 2009.
[15] L Wang, W Shen, H Xie, J Neelamkavil, and A Pardasani. Collaborative conceptual design - state of the art and future trends. 34, 2002.