Git-ing a Better Understanding of Commits

An Exploratory Study of Git Repository Commit Size Metrics

Drew Guarnera, Heather Michaud, Evan Purkhiser  
Department of Computer Science  
The University of Akron  
{dtg3,hmm34,emp36}@zips.uakron.edu

*Abstract*—Version control is an essential tool of modern software development. It provides the means to reverse detrimental changes with minimal effort, allows developers to work collaboratively on large scale applications, and can provide a detailed history of an application’s lifecycle. The historical archive of a system proves crucial in revealing valuable patterns, and the understanding of these evolutionary patterns can improve the discipline of software development. Many version control systems exist, however this work specifically focuses on the Git version control tool, a modern system that is rapidly gaining ground within the software development community. Since Git Write follows a ‘distributed version control’ model, any source code repository with a publicly accessible Git endpoint may be cloned by anyone using the Git tool. Changes made to each repository can live completely independent or may be pushed or pulled between clones. Due to this flexibility, Git truly sets itself apart from older and more established version control tools by offering users a truly customizable workflow.

This study is an attempt to improve upon existing work by examining the commit history for 13 Git repositories. This data is analyzed and compared to the foundational work to determine if use of the Git version control system changes the way developers make commits. Additionally, commit size metrics are calculated to investigate if oversized commits are prevalent during development, considering that more modern practices and version control tools such as Git emphasize on small commits.

Keywords—DVCS; Git; Source Control; Source Code Metrics

# Introduction

Throughout the course of software’s life cycle, any changes made to the project are recorded in the project’s version control system. When that version control system is Git, modifications are made to the developer’s local repository by committing the changes, and these local modifications are put into the central repository by pushing the changes. By exploiting the qualities of a version control system, the entire history of a project can be obtained through commit records.

Modern software practices encourage small commits to gradually improve the system and decrease the likelihood of bugs [citation?]. Git specifically motivates these practices with the commit-push workflow. Smaller changes can be committed, and the larger aggregation of changes can be pushed.

We feel that by studying repositories using modern version control system such as Git, we can gain a better understanding of common software development practices in terms of commit sizes. In previous studies [commit size distribution] [typical commit], commit size metrics were obtained from repositories and it was found that less than 15% of commits were considered large. In [typical commit] specifically, only repositories that used SVN were studied. Our ultimate goal is to determine if large commits are a frequent and problematic occurrence in repositories using Git, whose workflow exemplifies modern practices and stresses small commits. To address this goal, we examined three size metrics utilizing a tool we developed for traversing repository commit histories:

1. Number of lines of code (LOC) added, deleted, or modified in a commit
2. Number of files added, deleted, or modified in a commit
3. Number of hunks added, deleted, or modified in a commit

All metrics were obtained using the pygit2 [pygit2] library, which exposes the core functionality of Git. The aggregate files that have been modified in a commit contain the number of lines of code that have been modified at the lowest measured level of granularity. In between LOC and files measurements lie the number of hunks modified. A hunk is a group of lines that have been modified together within a file.

The study presented here examines 13 open source software projects. After collecting data for each project’s entire revision history, each calculated metric for lines, files, and hunks are separated into categories of extra-small, small, medium, large, and extra-large. These categories allow us to calculate trends over system and study relationships between categories and metrics by computing correlation coefficients.

The paper is organized as follows. Section 2 discusses related work in commit size studies, including the original study for which this paper replicates. Section 3 presents a background of Git, discusses how it is separated from other distributed version control systems and common workflows associated with Git. Section 4 presents design and implementation details of our python tool, srcstat, created to analyze source code size metrics regarding multiple repositories’ commit histories. Section 5 demonstrates an evaluation of the results gathered. Section 6 explores threats to the validity of the results. Section 7 describes future work, and Section 8 concludes.

# Related Works

Arafat and Riehle [commit size distribution] analyzed commits for over 9,000 open source projects and over 8 million commits using an analytics tool called Ohloh [ohloh]. Each commit was separated into three categories including single focused commits, aggregate team contributions, and repository refactorings. Single individual developer contributions deal with commit changes that ideally fix only one semantic issue, ranging from 1 to 100 SLOC in the commit size. Aggregate developer contributions consist of a set of commits that either a single developer or a team of developers made, similar to a Git push, ranging from 101 to 10,000 SLOC in the commit size. Component or repository refactoring or consolidation are over 10,000 SLOC in commit size and  often deals with integrating an entire library or branching a project. Categories were formed based on the size of the commit measured by SLOC. They found that single commits constitute 83% of their sample population. The authors suggested that the dominance of small commits is a result of modern open source software development processes, such as dictator and lieutenants workflow. For the larger commits, the authors proposed that such large commits were caused by copy and paste, initial check-in of existing projects, or merging.

Hindle, German, and Holt [study of large commits] created a taxonomy of large commits which expanded Swanson’s categories of changes [swanson], and manually classified the largest 1% of commits in nine open source projects. Large commits are often explained by merges, copyright changes, and license updates. The study found that the most frequently occurring large commits involved merges, feature additions, documentation, adding a module, initializing a module, code cleanup or refactoring, token replacement, bug fixes, and changes in the build or configuration settings. A theme of large commits was also considered for each project.

Hindle, German, Godfrey, and Holt [automatic classification of large changes] manually classify large commits from nine open source projects into categories that were an extension of Swanson’s categorization [swanson], then created a machine learning tool to automatically classify the large changes based on commit messages.

Abdulkareem Alali et. al. [typical commit] focused on the version histories of nine open source software projects that use SVN. Commit sizes were measured by LOC, hunks, and number of files modified using GNU diff. This allowed for varying levels of granularity with respect to the amount of changes made to the system, component, or module, while also respecting the cohesive properties of the system or module. Each size metric was separated into categories ranging from extra small to extra large. After all commits had been categorized, a linear correlation coefficient was calculated for lines and hunks, files and hunks, and files and lines. These correlations were also corresponded to the categories ranging from extra small to extra large. No significant correlation was found between line and file measures, but a substantial co-relationship was found between line and hunk measures. As opposed to SVN, our paper aims to replicate this study using Git. In addition, Abdulkareem Alali et. al also examined the vocabulary versus commit size, establishing the most frequently used words together based on the size of the commit.

# Background

Git is a distributed version control system (DVCS) offering speed and flexibility not found in other well-known version control systems. Originally developed in 2005 by Linus Torvalds for the primary purpose of managing the source code of the Linux kernel, Git has recently taken the software development world by storm. With the advent of social coding platforms such as GitHub, Bitbucket, and Gitorious, Git has seen widespread growth with over 6 million repositories hosted on GitHub alone.

## Distributed Version Control

Git sets itself apart from many other version control systems thanks to the fact that it offers a distributed approach to source control. While most well established version control systems follow a centralized model, such as SVN, where the source code exists in one central location and all changes to the source are made to that central repository, Git does not. Instead of “checking out” a local copy of a remote repositories working tree like with SVN, Git allows for a complete copy (known as a “clone” operation) of any given repository to be made. Each clone of a Git repository can be thought of as a completely independent repository that simply shares ancestral commits. When committing to a cloned repository the changes will only become available in that particular “clone” of the repository. Git does offers the ability to share these changes between repositories using the concepts of ‘pushing’ and ‘pulling’ changes between repositories that have previous commits in common (ancestral commits).

While most version control systems support the concept of branching Git makes this a first class feature of the tool and it becomes a core concept of the nature of Git. Git offers the ability to very quickly branch off of the current tip of a checked out branch, creating a new branch referencing that same commit. When making commits on a newly created branch these commits will only be available on the new branch while the original branch will be unchanged. By default, all repositories start with a ‘master’ branch.

Branching also plays a key part in the distributed nature of Git as any cloned repository may be thought of as a “remote branch” from which it was cloned (known as the “origin”). It is also possible to have a remote branch of any clone of the repository; Git makes no distinction between these remote branches. Changes between any branches can then be pushed and pulled between local branches, remote and local branches, or even between two remote branches.

## DVCS Workflows

Because of the distributed nature of Git, the user has the ability to completely customize their workflow of both their usage of Git, as well as the workflow for the entire project team. These workflows can become relevant in the study of commit metrics thus the following workflows have been taken into consideration:

#### Incremental Commit

A common way for new users to work with Git is to use the very basic workflow of continuously making changes to the source code, and incrementally committing their changes to the projects master branch as they make these changes. If working in a team environment, there will often only be one central Git repository that committers push their changes to. This mirrors the SVN way of working in that branching is rare and commits are made and pushed to a central location.

#### Reorganizing Commit Workflow

Because Git commits are made to the local repository it is possible to, at any given time, “go back in history” and reorder, remove, break up, or “squash” commits together. These features allow for a unique workflow in which the user can spend extra time organizing their commits after they have already been made, allowing for a cleanly organized history where each commit makes logical sense.

#### Branched-Merging Workflow

A common branching strategy with Git is to create a “topic branch” for each new feature to be added to the project. This is allows for work on a particular feature to be done independent of any work on the master branch and can offer a clean separation of workflow concerns. When changes are completed in a branch and are ready to be brought into the master branch of the project it is possible to create a “merge” commit. This is a special type of commit that has more than one parent commit reference. All changes made since the most recent ancestral commit of both parents will be reflected in this commit; however both lines of work are kept track of.

#### Branched-Squash Workflow

Similar to the Branched-Merge workflow, it’s also possible to squash multiple commits on a branch down into a single commit when merging it back into the mainline branch. This will discard entire branches commit history and merge the changes in as a single commit. This workflow is often preferred for projects where topic branches tend to be smaller and more concise features, as opposed to long running feature branches.

# SRCSTAT

## Approach

In order to properly replicate the initial study [typical commit], multiple Git repositories would need to be cloned to a local machine for the required analysis. The commits of each repository would then require analysis, and some of the repositories utilized in this study could contain anywhere from around ten thousand to several hundred thousand commits. To process the commits and obtain the required code change size metrics, a diff would need to be calculated between all consecutive pairs of commits starting with the initial commit and ending with the most recent commit to the repository, i.e.

git diff c1 c2, git diff c1 c2, etc.

The metrics of interest from each commit pair diff are the number of files modified, number of hunks, and number of lines modified between the commit. These values will be totaled separately per diff and grouped into the categories of extra small, small, medium, large, and extra-large based on their respective number of occurrences.

With the amount of data in need of analysis, tool assistance would be an absolute necessity in order to compile the metrics for this study. The initial approach taken toward tooling was to avoid reinventing the wheel and seek out an existing tool, and while there was no shortage of tools for software metrics, Cloc [cloc] and Ohcount [ohcount] initially stood out. Unfortunately while vetting the tools during the early testing and planning phase, it became clear that while Cloc and Ohcount would indeed provide software size metrics such as SLOC and LOC, these tools were unable to assist with providing metrics based on revision differences since the granularity was limited to the calculation of size metrics for an entire file or source code project directory.

When the search for off the shelf tooling proved unsuccessful, the Git command line client was then examined. This option was more promising as the Git command line tool has a very convenient option for its diff functionality called ‑‑shortstat. When combined with two revision identifiers the command provided a very clean single line diff response that detailed the number of files changed and the number of line additions and deletions. The only shortcoming of this approach was that finding the number of hunks in each commit was troublesome as that information is not included in the abbreviated diff report information. Using the standard diff option and parsing the full output of the diff report while not impossible, seemed inefficient and time consuming, so the search for a cleaner and more comprehensive solution lead to the libgit2 [libgit2] library (covered in the following section).

The initial tool while functional was quite basic. Test repositories were hard coded into the application, and upon execution, the repositories were cloned from GitHub [github]. Once a repository was downloaded, an array containing all of the commits, represented by cryptographic SHA hashes, was created and ordered from oldest to newest commit. These hashes were fed in pairs to a differencing function in the library which performed the internal Git differencing between the two revisions of the repository, and provided access to a data structure containing information on the files, hunks (consecutive blocks of code changes), and lines all altered between the two revision commits. This first endeavor proved to be a solid base for the project, but the performance of the approach proved inefficient. Upon closer analysis of the raw data being processed, it became clearer as to why the first attempted failed.

While a large amount of data was expected from the diff process, it was unclear during early testing of the tooling just how large some of the diffs would be. One particular repository, FreeBSD caused significant resource strain on the tool due to the sheer size of the diff object produced. As an example, one particular diff contained approximately seven thousand file modifications, consisting of nearly 250,000 line additions and 500,000 line deletions. Large commits such as this had serious ramifications on the tool’s performance, and ultimately exhausted all of the RAM in our two data collection systems, each equipped with eight gigabytes of RAM and crashed the OS (Ubuntu 12.04 64-bit and Arch Linux 3.12.1 64-bit). After observing this limitation first hand, adjustments were made in an effort to improve the efficiency and stability of the tool.

## Final Tool Design and Implementation

After experimentation with various tooling options, the authors’ arrived at a combination of libgit2 with Python for the core implementation of the final tool supplemented by output from the Git command client’s diff --shortstat option. Libgit2 is used to expose the core functionality of Git allowing for third party applications to gain programmatic access to the details of a Git repository through the use of its API. Since the library is utilized in production by companies such as GitHub and Microsoft [libgit2], this lent the library additional credibility and made the choice fairly obvious for the authors. While the library is written in C, there are bindings for fifteen additional languages [libgit2], one of which being the Python programming language. While this made the development language decision easier, the authors’ did consider various languages before choosing Python. In the end, it was thought that the due to the authors’ previous experience with Python development and the ease of quickly prototyping in the language that Python would prove to be a safe choice. Another nice feature of Python is the comprehensive standard library that provides many additional convenient features with no additional configuration or external libraries required. Once initial setup of the library and language bindings were complete, development with pygit2, libgit2 and Python proved to be was relatively painless.

The final iteration of our data collection tool resulted in a straightforward command line application capable of cloning a repository from GitHub and totaling and categorizing the size metric data for each commit pair diff comparison in the entire history of the repository. The tool supports command line options for processing an individual repo with the following command:

srcstat --repo [repo-uri]

Additionally, and entire list of repos can be processed by providing a text file consisting of a list of repository URLs with the following command:

srcstat --repolist [repo-list-file]

Once the tool has collected a repository, the pygit2 library is used to retrieve the entire set of cryptographic SHA hash that represent each individual commit, and the hashes are returned in a list sorted from oldest to most recent. These hashes are run through a python sub process instance of the git command line tool to run a git diff --shortstat. While the application incurs an execution time running this external process it allows for a significant reduction in the amount of data objects that need to be traversed to acquire the number of lines deleted and added per commit. While this data accounts for two of the three required code size metrics, collection of the hunks metric is a more complicated matter.

Collection of hunks requires the pygit2 diff function, which accepts the same two SHA hashes as the git diff --shortstat command. The result of the function is a complex data structure that allows for iteration over changed files, hunks, and lines of code. Since lines are already captured, our tool only need to traverse the changed files to find all hunks within each file and collect the total number of hunks per commit diff calculation. Once all the totals for any given diff are calculated, each metric for changed files, hunks and lines of code are then evaluated into extra small, small, medium, large, or extra-large categories based on their respective number of occurrences. When the entire repository is processed the end result is a file that contains a count the total number of modified files, hunks, and lines that fall into the aforementioned categories. It is this file and the values it contains that is used for the evaluation of our study. While this final revision of the tool was faster than the previous incarnations, it still struggled with the same memory consumption issues of its predecessor on the FreeBSD repository. With this in mind, the authors determined that a higher performance language such as C or C++ would most likely be necessary to provide the maximum efficiency necessary repositories with a combination of very large commits and a high overall commit history.

# Evaluation

## Setup

Utilizing the srcstat tool, data was collected from thirteen open source software projects with varying sizes and domains that were cloned from Git. Several repositories were included from a previous study [typical commit], such as gcc, Ruby, and Python. The aim of examining these repositories was to each repository’s commit history was traversed to obtain three size metrics:

1. Lines of code (LOC) added, deleted, or modified in a commit
2. Number of files added, deleted, or modified in a commit
3. Number of hunks added, deleted, or modified in a commit

Merge commits were also taken into account during the commit measures. Lines of code changes were calculated by taking the sum of line changes for each file changed. Hunks are sections of line changes that have been modified together within the file. For example, when lines 1-10 have been modified and lines 23-30 have been modified within the same file, two hunks exist of size 10 and 8 respectively. Each metric was separated into a category of sizes ranging from extra-small to extra-large, as defined by [typical commit] in a previous study. The representation of how each range was categorized is shown in Table 1.

| Metric | x-Small | Small | Medium | Large | x-Large |
| --- | --- | --- | --- | --- | --- |
| No. Files changed | 1-1 | 2-4 | 5-7 | 8-10 | 11-81880 |
| No. New lines | 0-5 | 6-46 | 47-106 | 107-166 | 167-54794 |
| No. Hunks | 0-1 | 2-8 | 9-17 | 18-26 | 27-58817 |

**Table 1. Range Categorizations for each commit size metric.**

## Examined Projects

Thirteen open source projects located in GitHub were examined for the purposes of this study. The projects are listed in Table 2 with their corresponding duration of project length, total number of commits throughout the project, total lines of code per repository, and the total number of files in the project.

| System | Duration | Total Commits | Total LOC | Total Files |
| --- | --- | --- | --- | --- |
| *Django Web Framework* | 8 Years | 16476 | 1021488 | 4939 |
| *Express Web Framework* | 4 Years | 4233 | 52060 | 347 |
| *GNU Compiler Collection (gcc)* | 25 Years | 127070 | 21523093 | 82630 |
| *GNU Image Manipulation Program (GIMP)* | 16 Years | 33588 | 4873393 | 5650 |
| *Apache Hadoop* | 5 Years | 8469 | 3025318 | 6219 |
| *jQuery JavaScript Library* | 8 Years | 5450 | 116980 | 235 |
| *libgit2* | 5 Years | 5915 | 247723 | 2504 |
| *Linux* | 8 Years | 413028 | 26422160 | 44984 |
| *Mono* | 13 Years | 94728 | 10237555 | 45707 |
| *Node.js* | 5 Years | 9397 | 2985235 | 9747 |
| *Python Language* | 23 Years | 41239 | 1880911 | 3971 |
| *Ruby Language* | 15 Years | 33789 | 2338318 | 4514 |
| *XBMC Media Center* | 4 Years | 25383 | 6559947 | 15701 |

Table 2. Thirteen open source projects analyzed by the srcstat tool.

Each repository is briefly described here in terms of its purpose and usage within the development community.

#### Django Web Framework

Django is an open source web applications framework for rapid development of database driven web applications using the Model View Controller (MVC) paradigm written in Python. Originally developed in 2005 Django has seen a steady rise in usage since.

#### Express Web Framework

Express is a relatively new web applications framework that utilizes the Node.js scripting language. Like Django, Express provides a way to rapidly create complex database driven MVC applications. Express is designed to be especially minimal in its "Don't do too much" approach and mostly handles HTTP logic. Development began just months before the initial Node.js release in January 2009.

#### GNU Compiler Collection (gcc)

The GNU Compiler Collection is a compiler suite produced for the GNU project. Gcc offers support for various programming languages. Gcc is widely popular, is available on almost all Linux distributions, and is the official compiler for the GNU operating system. Originally developed in 1987, gcc has played a large part in the growth of open source software.

#### GNU Image Manipulation Program (GIMP)

The GNU Image Manipulation Program is a cross platform open source photo and image manipulation tool that offers high end features commonly only available in commercial photo editing software. Development started in 1996 and like gcc this is also part of the GNU project.

#### Apache Hadoop

Hadoop is a open source framework developed by the Apache foundation in 2005 used for the processing and storage of data sets on a very large scale. Hadoop is split up into multiple modules, for the purpose of this project we will be focusing on the "Hadoop Common" module.

#### jQuery JavaScript Library

JQuery is a client side JavaScript library, primary concerned with providing a simple API interface to the Document Object Model, but also has a strong focus on normalizing browser quarks. JQuery originally began development in 2006 and has since become the most popular JavaScript library used today.

#### Libgit2

Libgit2 is a library which implements and exposes Git functionality through a C API. Libgit2 was originally developed in 2008 libgit2 is now used by the likes of GitHub, Beanstalk, and Microsoft. Language bindings are also offered for various other languages.

#### Linux

Originally developed by Linus Torvalds in 1995 as a spinoff of the Minx kernel, the Linux Kernel provides the foundational backbone for many open source and commercial operating system distributions. Its proven effectiveness and stability has earned Linux its place as one the most popular operating systems around the world.

#### Mono

Mono is an open source compiler that is compatible with C# and the dot NET framework. It also includes the common language runtime required to execute programs compiled with mono. Mono originally began development in 2004 by Novell, but has since been taken over by Xamarin.

#### Node.js

Node.js is relatively new software platform used for creating highly scalable network applications. The platform utilizes the JavaScript scripting language and uses an event based non-blocking event loop. Development only recently begun in 2009, but it has since gained a lot of traction due to the need for scalable web applications.

#### Python Language

Python is a powerful dynamic programming language with a syntax designed to be clear and readable. Backed by an extensive library, including a wide array of functionally provided "out of the box" python is a language that has proven to be quite applicable for rapid application development and prototyping.

#### Ruby Language

Publicly released in 1995, Ruby is functional programming language that strives to provide an clean object oriented language that is easy to comprehend and powerful in application. When combined with the Rails web framework, Ruby can be used to code web services and applications.

#### XBMC Media Center

Created in 2003 XBMC is a cross platform and open source entertainment management and presentation application. Designed to provide a simple and attractive playback for both audio and video multimedia, XBMC is frequently used as a HTPC (Home Theatre Personal Computer) interface. XBMC has received the support of more than 450 software developers worldwide, and is available in more than 65 languages.

Fig 1. All File lines hunks, BETTER DESCRIPTION NEEDED

**Fig 2. gcc File line hunks, BETTER DESCRIPTION NEEDED**

## Quintessential Commit Sizes

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Number of Files** | | | **Number of New Lines** | | | **Number of Hunks** | | |
| **Categories** | **Range** | **Freq.** | **Ratio** | **Range** | **Freq.** | **Ratio** | **Range** | **Freq.** | **Ratio** |
| **x-Small** | 1-1 | 348790 | 43% | 0-5 | 174793 | 21% | 0-1 | 195791 | 31% |
| **Small** | 2-4 | 308401 | 38% | 6-46 | 358576 | 44% | 2-8 | 421850 | 37% |
| **Median** | 5-7 | 56167 | 7% | 47-106 | 100408 | 12% | 9-17 | 80899 | 9% |
| **Large** | 8-10 | 20049 | 2% | 107-166 | 39095 | 5% | 18-26 | 26593 | 3% |
| **x-Large** | 11-19543 | 81974 | 10% | 167-8043 | 145893 | 18% | 27-11095 | 93632 | 20% |

Table 2. Aggregate commits from all 13 projects examined, categorized by size for each of the three commit size metrics.

Here we determine into which size range most commits are categorized. This will demonstrate the most frequent size of changes that have been committed. The ratio of commit sizes for the gcc project are shown in Figure 3, and the ratio of the average commit size for each open source project and category are shown in Figure 4. As seen in Figure 3, 64% of commits have been categorized as small or extra small based on the lines of code added, deleted, or modified. 78% have been categorized as small or extra small based on the number of files changed, and 76% of commits were classified as small or extra small based on the number of hunks modified.

The average commit size for each of the 13 open source project separated by size categories are shown in Figure 4. Aggregating all open source projects studied, 23% of the lines of code modified, 12% of files modified, and 23% of hunks modified were classified as large or extra-large. With the exception of the files modified, all other commit size metrics examined contained over 20% of larger commits. Based on these results, large commits are still a common occurrence, even given the modern software development practices when utilizing Git.

## Correlation between Characteristics

In order to determine relationships between characteristics, the correlation coefficient is calculated between each metric. The correlation coefficient , ranging [-1, 1], attempts to calculate if a linear co-relationship exists between variables. A negative value for r implies that as x increases, y decreases. A positive value for r implies that as x increases, y decreases. The closer the coefficient is to +1 or -1, the stronger the relationship. When , no linear relationship exists between variables. [applied multivariate statistical analysis]. A strong correlation is determined with r-values surpassing 0.8, while weak correlations are considered to have r-values less than 0.5. The equation used to calculate the correlation coefficient is as follows

The coefficient of determination, , is a statistical calculation that indicates how well the relationship between variables fits into a linear regression model, ranging [0, 1]. It measures the strength of the correlation between variables. For example, if , then , meaning that 67.4% of the total variation between x and y can be explained by the linear relationship defined by . Thus the remaining 32.6% of the variation remains unexplained by a linear model.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Number of Files** | | | **Number of New Lines** | | | **Number of Hunks** | | |
| **Categories** | **Range** | **Freq.** | **Ratio** | **Range** | **Freq.** | **Ratio** | **Range** | **Freq.** | **Ratio** |
| **x-Small** | 1-1 | 16324 | 13% | 0-5 | 25339 | 20% | 0-1 | 17462 | 14% |
| **Small** | 2-4 | 65300 | 51% | 6-46 | 74267 | 58% | 2-8 | 79091 | 62% |
| **Median** | 5-7 | 18613 | 15% | 47-106 | 14532 | 11% | 9-17 | 14226 | 11% |
| **Large** | 8-10 | 7290 | 6% | 107-166 | 4809 | 4% | 18-26 | 5196 | 4% |
| **x-Large** | 11-19543 | 19543 | 15% | 167-8043 | 8043 | 6% | 27-11095 | 11095 | 9% |

Table 3. *gcc* commits categorized by size for each of the three commit size metrics. The duration of the commits was over 25 years.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **files × lines** | **files × hunks** | **hunks × lines** |
| ***Django*** | ***r*** | 0.821 | 0.891 | 0.986 |
| ***r²*** | 0.674 | 0.794 | 0.973 |
| ***Express*** | ***r*** | 0.885 | 0.947 | 0.985 |
| ***r²*** | 0.782 | 0.898 | 0.969 |
| ***gcc*** | ***r*** | 0.963 | 0.988 | 0.989 |
| ***r²*** | 0.928 | 0.977 | 0.979 |
| ***GIMP*** | ***r*** | 0.694 | 0.907 | 0.903 |
| ***r²*** | 0.481 | 0.823 | 0.816 |
| ***Hadoop*** | ***r*** | 0.806 | 0.994 | 0.809 |
| ***r²*** | 0.650 | 0.989 | 0.655 |
| ***jQuery*** | ***r*** | 0.853 | 0.876 | 0.995 |
| ***r²*** | 0.728 | 0.767 | 0.991 |
| ***Libgit2*** | ***r*** | 0.893 | 0.873 | 0.941 |
| ***r²*** | 0.798 | 0.762 | 0.886 |
| ***Linux*** | ***r*** | 0.580 | 0.606 | 0.982 |
| ***r²*** | 0.336 | 0.367 | 0.965 |
| ***Mono*** | ***r*** | 0.966 | 0.991 | 0.986 |
| ***r²*** | 0.934 | 0.982 | 0.971 |
| ***Node*** | ***r*** | 0.721 | 0.836 | 0.949 |
| ***r²*** | 0.520 | 0.699 | 0.900 |
| ***Python*** | ***r*** | 0.648 | 0.762 | 0.983 |
| ***r²*** | 0.421 | 0.580 | 0.965 |
| ***Ruby*** | ***r*** | 0.976 | 0.989 | 0.997 |
| ***r²*** | 0.953 | 0.978 | 0.994 |
| ***XBMC*** | ***r*** | 0.840 | 0.848 | 0.988 |
| ***r²*** | 0.706 | 0.720 | 0.976 |

Table 7. Correlation coefficient, *r*, and coefficient of determination, *r²*, as calculated for all projects.

The correlation coefficient and coefficient of determination for gcc are shown in Table 4. As can be seen, all correlation coefficients are above 0.96. This means that all relationships between files and lines, files and hunks, and hunks and lines are strongly linearly correlated. Over 92% of the variants can be explained by this linear correlation.

All thirteen projects showed a strong correlation between hunks and lines size metrics. Only Hadoop contained a correlation coefficient under 0.9, holding a correlation of 0.809. All other repositories had r-values exceeding 0.9, indicating a very strong positive linear correlation between hunks and lines modified in a commit.

Most projects had a strong correlation between files and hunks. Only Linux and Python had moderate correlations, containing 0.606 and 0.762 respectively. The inconsistency with Linux may be due to the nature of the project. Linux is a long running and stable project, which is based deeply on providing consistent and reliable functionality. With this in mind, changes would most likely be limited to certain sections in a file. Linux also has very rigid development restrictions that the maintainers placed on contributors to keep commits small and focused. Additionally, the moderate correlation can be described in terms of commit distributions. In terms of the number of hunks modified per commit, 27% and 48% of commits were classified as extra small or small, respectively. However, the number of commits classified as such when measured by the number of files changed is nearly reversed.

The minor inconsistency with Python’s moderate correlation between files and hunks can be explained in multiple aspects. One such explanation is that Python has under 4,000 files and, according to our previous measurements, only 3% of all commits made to the project were considered to be large or extra-large based on the number of files modified. In addition, only 5% of all commits made to the project were categorized as large or extra-large based on the number of hunks modified in a commit. Another explanation follows. Manny of the modules in Python’s standard library have a long-standing code base which require minimal architectural changes. This leads to the assumption that only small maintenance related changes are the focus of a majority of the commits. An alternative interpretation of the results could be associated with a work flow that consists of development policies to distill large changes or maintenance work into a batch of smaller commits. This would directly correlate with Git’s ability to easily reorganize commits as mentioned previously in Section 2.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Number of Files** | | | **Number of New Lines** | | | **Number of Hunks** | | |
| **Categories** | **Range** | **Freq.** | **Ratio** | **Range** | **Freq.** | **Ratio** | **Range** | **Freq.** | **Ratio** |
| **x-Small** | 1-1 | 29291 | 71% | 0-5 | 13126 | 32% | 0-1 | 16930 | 41% |
| **Small** | 2-4 | 9304 | 23% | 6-46 | 17916 | 43% | 2-8 | 19314 | 47% |
| **Median** | 5-7 | 1264 | 3% | 47-106 | 4249 | 10% | 9-17 | 2765 | 7% |
| **Large** | 8-10 | 404 | 1% | 107-166 | 1756 | 4% | 18-26 | 845 | 2% |
| **x-Large** | 11-19543 | 895 | 2% | 167-8043 | 4192 | 10% | 27-11095 | 1385 | 3% |

Table 4. *Python* commits categorized by size for each of the three commit size metrics.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Number of Files** | | | **Number of New Lines** | | | **Number of Hunks** | | |
| **Categories** | **Range** | **Freq.** | **Ratio** | **Range** | **Freq.** | **Ratio** | **Range** | **Freq.** | **Ratio** |
| **x-Small** | 1-1 | 224237 | 54% | 0-5 | 101299 | 25% | 0-1 | 112204 | 27% |
| **Small** | 2-4 | 102942 | 25% | 6-46 | 163916 | 40% | 2-8 | 190574 | 46% |
| **Median** | 5-7 | 22553 | 5% | 47-106 | 47802 | 12% | 9-17 | 39474 | 10% |
| **Large** | 8-10 | 7788 | 2% | 107-166 | 18131 | 4% | 18-26 | 11959 | 3% |
| **x-Large** | 11-19543 | 54794 | 13% | 167-8043 | 81880 | 20% | 27-11095 | 58817 | 14% |

Table 5. *Linux* commits categorized by size for each of the three commit size metrics.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Number of Files** | | | **Number of New Lines** | | | **Number of Hunks** | | |
| **Categories** | **Range** | **Freq.** | **Ratio** | **Range** | **Freq.** | **Ratio** | **Range** | **Freq.** | **Ratio** |
| **x-Small** | 1-1 | 4589 | 49% | 0-5 | 2285 | 24% | 0-1 | 3015 | 32% |
| **Small** | 2-4 | 3031 | 33% | 6-46 | 3476 | 37% | 2-8 | 4391 | 47% |
| **Median** | 5-7 | 470 | 5% | 47-106 | 1467 | 16% | 9-17 | 735 | 8% |
| **Large** | 8-10 | 203 | 2% | 107-166 | 445 | 5% | 18-26 | 243 | 3% |
| **x-Large** | 11-19543 | 1025 | 11% | 167-8043 | 1724 | 18% | 27-11095 | 1013 | 11% |

Table 6. *Node* commits categorized by size for each of the three commit size metrics.

Django, Express, gcc, Hadoop, jQuery, libgit2, and Mono had a strong correlation between files and lines. GIMP, Linux, Node, and Python had *r* values of 0.684, 0.58, 0.721, and 0.684 respectively, indicating moderate correlations between files and lines. GIMP’s commit distributions could have skewed the results of the correlation calculations because only 16% of the number of files modified in a commit were categorized as large or extra-large, where as 38% of the number of lines of code modified in a commit were categorized thusly. The focus of the GIMP application is centered around user interaction and the ability to edit images and graphical projects. With user satisfaction and maintaining a rich feature as primary objectives, it stands to reason that additional feature additions and architectural modifications would be more frequent in order to maintain relevance among the application’s user base. These changes translate to multiple line modifications mostly centralized to, or at least built on top of, a given module’s file containing the functionality.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Number of Files** | | | **Number of New Lines** | | | **Number of Hunks** | | |
| **Categories** | **Range** | **Freq.** | **Ratio** | **Range** | **Freq.** | **Ratio** | **Range** | **Freq.** | **Ratio** |
| **x-Small** | 1-1 | 7702 | 23% | 0-5 | 3967 | 12% | 0-1 | 4327 | 13% |
| **Small** | 2-4 | 16754 | 50% | 6-46 | 12523 | 37% | 2-8 | 15184 | 45% |
| **Median** | 5-7 | 3297 | 10% | 47-106 | 4487 | 13% | 9-17 | 4929 | 15% |
| **Large** | 8-10 | 1704 | 5% | 107-166 | 2279 | 7% | 18-26 | 2172 | 6% |
| **x-Large** | 11-19543 | 3772 | 11% | 167-8043 | 10332 | 31% | 27-11095 | 6976 | 21% |

Table 8. *GIMP* commits categorized by size for each of the three commit size metrics.

In addition, the size distribution of Node’s commits varied greatly. Under 25% of the commits were considered extra-small based on LOC modifications, where as nearly 50% of commits were categorized as extra-small based on the number of file modifications. Oddities in the correlation between files and lines of Node could also be associated with the fact that Node has yet to reach its 1.0 milestone release, and is still under heavy development. This could make pattern determination difficult due to the potential volatility of the project.

Similarly as previously speculated, the inconsistencies identified with Python’s correlation between file and line sizes can be explained by the commit size distributions in addition to workflow patterns. The number of commits to Python that were classified as extra-small and small were 71% and 23% respectively when measuring based on files modified. However, only 32% and 43% of commits were classified thusly when measuring based on lines modified.

In Linux, 54% and 25% of the commits are classified as extra-small or small, respectively, based on the number of files modified, where as almost reversely, 25% and 40% of the commits were categorized as extra-small or small based on the number of lines modified in a commit. These differences in categorizing commit sizes based on the size metric would skew the correlation coefficient.

All projects have a moderate to very strong positive correlation between all characteristics examined during the study. The number of lines of code changed and the number of hunks modified had the strongest correlation across projects. The number of files and hunks modified had the next to strongest correlation, while the number of files and lines of code changed per commit had the least strongest but still moderate to strong correlations between characteristics.

# Threats to Validity

Certain threats to validity include human error during development and data collection. In addition, the number of commits per project varied slightly than when the Alil paper [typical commit] was published due to new releases and bug fixes since publication. The additional data keeps the study up to date, but could skew the results to vary from those found by Alil et al Our tool is also missing various commits from thrown exceptions rooted in pygit2’s library calls. The exceptions are ignored and the commit is skipped over, though limited data is lost.

An additional threat to validity is merged branches. Some of the repositories examined in the original paper [typical commit], such as gcc, Ruby, and Python, were originally developed using SVN. Specifically, *gcc* still uses SVN and the repository located on GitHub is only a mirror to the changes presently being made to SVN. However, these repositories were also found in GitHub, meaning that the project was moved over to Git at a later date. Thus these projects did not fully take advantage of the Git-specific utilities. For example, the projects developed using the SVN version control system commit changes use the SVN commit. Projects developed using Git follow the commit-push workflow, where a developer’s changes are committed to his local repository, then the aggregate of all commits are pushed to the central repository. In this way, Git encourages smaller changes which can then be pushed as a larger change, whereas SVN changes are possibly equivalent to the size of Git pushes. Therefore, those projects which were originally developed using SVN and later merged to Git will resemble development patterns and practices that closer resemble SVN styles as opposed to Git styles, which is the focus of this paper.

In addition, the inter quartile ranges that were used to calculate the size ranges from extra small to extra large were calculated by [typical commit] on *gcc* at the time of the paper. This study used those size ranges, though the study was performed on multiple repositories whose data points varied, and even the *gcc* repository is over twice as large as the original repository used in the previous study. This remains as future work.

# Future Work

Future work involves making an improved version of our tool that is quicker, more memory efficient, and capable of performing similar and additional commit size metric calculations on a large number of repositories.

Given the results of the study, future work involves developing an automated commit splicing tool that will recognize a large commit based on various size metrics and splice it into one or more smaller, logically structured commits. Such a tool would ameliorate the software development process, especially given our findings.

The previous study [typical commit] used inter quartile ranges to calculate how the commit size data should be distributed. These ranges were found by calculating seven regions for extreme outliers downward, mid outliers downward, non-outliers downward, in the box, non-outliers upward, mid outliers upward, and extreme outliers upward. When shown in a box and whisper plot, these ranges constitute the sizes ranging from extra small to extra large. In future work, we intend to calculate the inter quartile ranges for each of the repositories studied.

# Conclusion

The purpose of this study was to determine if large commits are a prevalent issue in modern software development practices. We believe that by studying repositories using modern version control system such as Git, our understanding of common software development practices in terms of commit sizes measured in terms of files, lines, and hunks can be improved. A previous study conducted in 2008 [typical commit] studied nine open source projects using SVN to determine the frequency of varying commit sizes and attempted to find a correlation between various forms of size metrics. In this work, Alali et al concluded that less than fifteen percent of commits were considered to be large. Our paper differentiates itself from the original work by replicating that study instead with a larger dataset and focusing on projects that utilize Git as the primary version control system as opposed to the more classical and well-established SVN.

The choice of version control system is a crucial element of the software development process. Using SVN allows developers to commit their changes to the project while the utilization of Git supports a commit-push workflow. This involves a developer making changes and then committing the changes to his local repository after which multiple commits can be pushed to the central repository. This paper examines the commit sizes of thirteen open source projects located in GitHub. Three size metrics were gathered including number of files, lines of code, and hunks added, deleted, or modified in a commit. These three size metrics were broken into five categories of extra small, small, medium, large, and extra large. The correlation coefficient between the different metrics and the coefficient of determination was calculated to determine if there was a co-relation between metrics, and how strongly the relationship could be described by a linear model.

Ultimately it was found that 12% to 23% of commits were classified as large or extra-large, depending on the size metric that was used. This leads us to conclude that large commits are still a prevalent issue in modern software development, even given that the workflow of Git encourages small commits. All size metrics were found to be moderately to be strongly correlated. Hunks and lines of code modified had the strongest correlation of the three measurements, as every one of the thirteen repositories studied had strongly correlated hunks and lines. Files and hunks followed closely, with all but two repositories having strongly correlated values, and the remaining projects having moderately correlated values. Files and lines did moderately well, with all but four projects having strongly correlated values. The remaining projects had moderately correlated values. Our final observation notes the frequency of large commits in modern practices.

##### References

1. Cloc, http://cloc.sourceforge.net/. 12/9/13
2. Ohcount, https://github.com/blackducksw/ohcount. 12/9/13
3. libgit2, http://libgit2.github.com. 12/9/13
4. Python, http://www.python.org. 12/9/13
5. pygit2, http://www.pygit2.org. 12/9/13
6. GitHub, https://github.com. 12/9/13

##### Apendix

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Number of Files** | | | **Number of New Lines** | | | **Number of Hunks** | | |
| **Categories** | **Range** | **Freq.** | **Ratio** | **Range** | **Freq.** | **Ratio** | **Range** | **Freq.** | **Ratio** |
| **x-Small** | 1-1 | 8427 | 52% | 0-5 | 4885 | 30% | 0-1 | 5956 | 36% |
| **Small** | 2-4 | 5517 | 34% | 6-46 | 7056 | 43% | 2-8 | 7550 | 46% |
| **Median** | 5-7 | 992 | 6% | 47-106 | 1821 | 11% | 9-17 | 1383 | 8% |
| **Large** | 8-10 | 368 | 2% | 107-166 | 666 | 4% | 18-26 | 427 | 3% |
| **x-Large** | 11-19543 | 990 | 6% | 167-8043 | 2048 | 12% | 27-11095 | 1160 | 7% |

Table 9. *Django* commits categorized by size for each of the three commit size metrics.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Number of Files** | | | **Number of New Lines** | | | **Number of Hunks** | | |
| **Categories** | **Range** | **Freq.** | **Ratio** | **Range** | **Freq.** | **Ratio** | **Range** | **Freq.** | **Ratio** |
| **x-Small** | 1-1 | 19458 | 21% | 0-5 | 12781 | 13% | 0-1 | 13808 | 15% |
| **Small** | 2-4 | 59577 | 63% | 6-46 | 49515 | 52% | 2-8 | 62259 | 66% |
| **Median** | 5-7 | 7886 | 8% | 47-106 | 13427 | 14% | 9-17 | 10176 | 11% |
| **Large** | 8-10 | 2768 | 3% | 107-166 | 5165 | 5% | 18-26 | 3254 | 3% |
| **x-Large** | 11-19543 | 4771 | 5% | 167-8043 | 13840 | 15% | 27-11095 | 5231 | 6% |

Table 10. *Mono* commits categorized by size for each of the three commit size metrics.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Number of Files** | | | **Number of New Lines** | | | **Number of Hunks** | | |
| **Categories** | **Range** | **Freq.** | **Ratio** | **Range** | **Freq.** | **Ratio** | **Range** | **Freq.** | **Ratio** |
| **x-Small** | 1-1 | 2093 | 52% | 0-5 | 1411 | 33% | 0-1 | 1795 | 42% |
| **Small** | 2-4 | 1540 | 39% | 6-46 | 2038 | 48% | 2-8 | 2140 | 51% |
| **Median** | 5-7 | 197 | 5% | 47-106 | 448 | 11% | 9-17 | 195 | 5% |
| **Large** | 8-10 | 77 | 2% | 107-166 | 106 | 3% | 18-26 | 41 | 1% |
| **x-Large** | 11-19543 | 90 | 2% | 167-8043 | 230 | 5% | 27-11095 | 62 | 1% |

Table 11. *Express* commits categorized by size for each of the three commit size metrics.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Number of Files** | | | **Number of New Lines** | | | **Number of Hunks** | | |
| **Categories** | **Range** | **Freq.** | **Ratio** | **Range** | **Freq.** | **Ratio** | **Range** | **Freq.** | **Ratio** |
| **x-Small** | 1-1 | 12044 | 48% | 0-5 | 7844 | 31% | 0-1 | 8461 | 33% |
| **Small** | 2-4 | 6809 | 27% | 6-46 | 9355 | 37% | 2-8 | 10127 | 40% |
| **Median** | 5-7 | 1349 | 5% | 47-106 | 2224 | 9% | 9-17 | 2062 | 8% |
| **Large** | 8-10 | 549 | 2% | 107-166 | 833 | 3% | 18-26 | 711 | 3% |
| **x-Large** | 11-19543 | 4162 | 17% | 167-8043 | 5127 | 20% | 27-11095 | 4022 | 16% |

Table 12. *XBMC* commits categorized by size for each of the three commit size metrics.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Number of Files** | | | **Number of New Lines** | | | **Number of Hunks** | | |
| **Categories** | **Range** | **Freq.** | **Ratio** | **Range** | **Freq.** | **Ratio** | **Range** | **Freq.** | **Ratio** |
| **x-Small** | 1-1 | 675 | 8% | 0-5 | 912 | 11% | 0-1 | 468 | 6% |
| **Small** | 2-4 | 4644 | 55% | 6-46 | 2975 | 35% | 2-8 | 4527 | 53% |
| **Median** | 5-7 | 1150 | 14% | 47-106 | 1242 | 15% | 9-17 | 1417 | 17% |
| **Large** | 8-10 | 513 | 6% | 107-166 | 669 | 8% | 18-26 | 580 | 7% |
| **x-Large** | 11-19543 | 1429 | 17% | 167-8043 | 2671 | 32% | 27-11095 | 1477 | 17% |

Table 13. *Hadoop* commits categorized by size for each of the three commit size metrics.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Number of Files** | | | **Number of New Lines** | | | **Number of Hunks** | | |
| **Categories** | **Range** | **Freq.** | **Ratio** | **Range** | **Freq.** | **Ratio** | **Range** | **Freq.** | **Ratio** |
| **x-Small** | 1-1 | 9633 | 29% | 0-5 | 6470 | 19% | 0-1 | 7465 | 22% |
| **Small** | 2-4 | 20606 | 62% | 6-46 | 20272 | 60% | 2-8 | 21661 | 64% |
| **Median** | 5-7 | 1804 | 5% | 47-106 | 3394 | 10% | 9-17 | 2587 | 8% |
| **Large** | 8-10 | 560 | 2% | 107-166 | 1199 | 4% | 18-26 | 806 | 2% |
| **x-Large** | 11-19543 | 854 | 3% | 167-8043 | 2454 | 7% | 27-11095 | 1270 | 4% |

Table 14. *Ruby* commits categorized by size for each of the three commit size metrics.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Number of Files** | | | **Number of New Lines** | | | **Number of Hunks** | | |
| **Categories** | **Range** | **Freq.** | **Ratio** | **Range** | **Freq.** | **Ratio** | **Range** | **Freq.** | **Ratio** |
| **x-Small** | 1-1 | 3167 | 58% | 0-5 | 1887 | 35% | 0-1 | 2160 | 40% |
| **Small** | 2-4 | 1756 | 32% | 6-46 | 2315 | 42% | 2-8 | 2576 | 47% |
| **Median** | 5-7 | 161 | 3% | 47-106 | 518 | 10% | 9-17 | 317 | 6% |
| **Large** | 8-10 | 81 | 1% | 107-166 | 184 | 3% | 18-26 | 113 | 2% |
| **x-Large** | 11-19543 | 249 | 5% | 167-8043 | 546 | 10% | 27-11095 | 284 | 5% |

Table 15. *jQuery* commits categorized by size for each of the three commit size metrics.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Number of Files** | | | **Number of New Lines** | | | **Number of Hunks** | | |
| **Categories** | **Range** | **Freq.** | **Ratio** | **Range** | **Freq.** | **Ratio** | **Range** | **Freq.** | **Ratio** |
| **x-Small** | 1-1 | 2135 | 39% | 0-5 | 1602 | 27% | 0-1 | 1740 | 29% |
| **Small** | 2-4 | 1654 | 30% | 6-46 | 1919 | 32% | 2-8 | 2456 | 42% |
| **Median** | 5-7 | 512 | 9% | 47-106 | 716 | 12% | 9-17 | 633 | 11% |
| **Large** | 8-10 | 225 | 4% | 107-166 | 372 | 6% | 18-26 | 246 | 4% |
| **x-Large** | 11-19543 | 900 | 17% | 167-8043 | 1306 | 22% | 27-11095 | 840 | 14% |

Table 16. *libgit2* commits categorized by size for each of the three commit size metrics.