How Do Centralized and Distributed Version Control Systems Impact Software Changes?

Caius Brindescu, Mihai Codoban, Sergii Shmarkatiuk, Danny Dig
School of EECS, Oregon State University
Corvallis, OR, USA
{brindesc,codobanm,shmarkas,digd}@eecs.oregonstate.edu

ABSTRACT

Distributed Version Control Systems (DVCS) have seen an increase in popularity relative to traditional Centralized Version Control Systems (CVCS). Yet we know little on whether developers are benefitting from the extra power of DVCS. Without such knowledge, researchers, developers, tool builders, and team managers are in the danger of making wrong assumptions.

In this paper we present the first in-depth, large scale empirical study that looks at the influence of DVCS on the practice of splitting, grouping, and committing changes. We recruited 820 participants for a survey that sheds light into the practice of using DVCS. We also analyzed 409M lines of code changed by 358300 commits, made by 5890 developers, in 132 repositories containing a total of 73M LOC. Using this data, we uncovered some interesting facts. For example, (i) commits made in distributed repositories were 32% smaller than the centralized ones, (ii) developers split commits more often in DVCS, and (iii) DVCS commits are more likely to have references to issue tracking labels.

Categories and Subject Descriptors: D.2.7 [Software Engineering]: Distribution, Maintenance, and Enhancement

General Terms: Measurement, Experimentation

Keywords: Version Control, Software Change, Distributed Version Control, Centralized Version Control

1. INTRODUCTION

Distributed Version Control Systems (DVCS) like Git [2] or Mercurial [6] are widely used today. Over the last couple of years GitHub [4], which is the most popular repository hosting service for Git projects, has taken the open source community by storm [19]. At the end of 2012, GitHub hosted over 4.6M repositories. Compare this with the previous paradigm, CVCS, epitomized by SVN [9] and CVS [1]. SourceForge [7], the primary repository hosting service for SVN had about 300K repositories by the end of 2012. Also, our own survey of 820 developers shows that 65% use DVCS and 35% use CVCS.

DVCS brings a whole set of novel capabilities. Using DVCS, developers (i) can work in isolation on local copies of the repositories enabling them to work offline while still retaining full project history, (ii) they can cheaply create and merge branches, and (iii) they can commit individual changed lines in a file, as opposed to being forced to commit a whole file like in CVCS.

Are developers truly taking advantage of these DVCS features or are they simply paying the steep learning price without benefiting from them? Despite the large scale adoption of DVCS, we know little about the state of the practice in using this new paradigm. Without such knowledge, developers and managers are left in the dark when deciding whether it is worth to invest time and effort to transition to these new tools. Also, researchers are in danger of making errors when mining repositories, due to confounding effects imposed by DVCS. Finally, tool builders can build the wrong tools if they are not aware of developers' habits.

In this paper we present the first large-scale study that answers in-depth questions about the extent to which DVCS influences the practice of managing changes. To this end, we designed and launched a survey. We recruited 820 participants, 85% of them being developers from industry. 56% have ten or more years of programming experience. 51% work in teams larger than 6 developers.

To get further insights into how DVCS affects code changes, also we analyzed 409M lines of code changes from 358300 commits, made by 5890 developers, in 132 repositories containing a total of 73M LOC. Our corpus contains both pure and hybrid repositories. Pure repositories use the same VCS throughout their lifecycle. Hybrid repositories started in the centralized paradigm and switched to the distributed paradigm. The hybrid repositories can reveal if changing the version control system influences developers' practices.

For the centralized paradigm we chose SVN as the best representative. For the distributed paradigm we chose Git.

Using the data from our survey and our mining of repositories, we answer 12 research questions organized in three overarching themes:

Theme 1: How does the VCS type affect developers' behavior?

RQ 1: *Does the type of VCS affect the size of commits?*

RQ 2: *Do developers split their commits into logical units of change? How do they do it?*

RQ 3: *How often and why do developers squash their commits?*

ICSE'14, May 31 – June 7, 2014, Hyderabad, India
Copyright 2014 ACM 978-1-4503-2756-5/14/05...$15.00
http://dx.doi.org/10.1145/2568225.2568322

RQ 4: *Why do developers prefer one Version Control System over another?*

RQ 5: *Does the VCS influence the frequency with which developers commit?*

We found that developers' behavior is influenced by the VCS type. When using DVCS, developers make commits 32% smaller and they organize their changes in several commits. Depending on the VCS type, the reasons why developers find the commit process more natural are different.

Theme 2: How does the team size affect VCS usage?

RQ 6: *Does team size affect the choice of VCS?*

RQ 7: *Are larger teams more likely to use Issue Tracking Systems (ITS)?*

RQ 8: *Does team size affect the size of commits?*

RQ 9: *Does team size influence commit squashing?*

Teams of all sizes prefer using DVCS. The team size does not influence the size of commits. Most teams include issue tracking labels in their commits.

Theme 3: How does the VCS type affect the development process?

RQ 10: *Does the type of VCS influence the presence and the number of issue tracking labels (ITL)?*

RQ 11: *Is there a correlation between the number of issue tracking labels in the commit message and the commit size?*

RQ 12: *How does the size of commits vary in time?*

We found that developers using DVCS include issue tracking labels more often in commit messages. Also, the commit size decreases as the project matures.

Based on these findings, we propose several actionable implications for four audiences. *Researchers* can better align their research questions with the type of repositories they mine. For example, for questions that rely on a discrete and precise software changes (e.g., bug prediction etc.) they should mine distributed repositories. *Developers* can give more precise meaning to their changes when they use DVCS. *Tool builders* can further build up on the strengths provided by DVCS such as the ability to better group changes and express their intent. *Managers* can make more informed decisions when choosing tools for their projects.

This paper makes the following contributions:

1. **Research Questions.** We designed and answered 12 novel research questions to understand the extent in which DVCS help developers manage software changes.
2. **Survey.** We designed and launched a survey to provide insights into the practice of using DVCS. We recruited 820 participants.
3. **Mining repositories.** We developed tools to collect metrics and analyze centralized and distributed repositories. We applied these tools on 132 repositories.
4. **Implications.** We present implications of our findings from the perspective of four audiences: *researchers*, *developers*, *tool builders*, and *team managers*.

The tools, summary of survey responses, and corpus are publicly available at:

`http://cope.eecs.oregonstate.edu/VCStudy`

2. EXPERIMENTAL SETUP

In this section we describe the two sources of data we used to answer our research questions.

2.1 Survey

We conducted a survey where we asked 20 questions about developer commit practices. 820 respondents answered our survey. The participants are developers recruited by promoting the survey on social media channels specific to the development community, i.e., Twitter and Google+ feeds that are mainly read by developers.

Table 1 shows the demographics of the respondents. Most are experienced developers working on industrial projects. The data shows that Git is widely used by developers (52%), followed by SVN (20%).

Table 1: Demographics of survey respondents

(a) Programming experience (years)

< 2	2 - 5	5 - 10	10 - 15	15 - 20	> 20
1.83%	11.10%	30.49%	30.61%	13.90%	12.07%

(b) Project type

Proprietary software	Open source software	Research project	Personal project	Other
85.09%	6.97%	4.64%	3.06%	0.24%

(c) Team size

1	2 - 5	6 - 10	11 - 25	26 - 100	> 100
5.87%	42.30%	23.72%	15.65%	8.19%	4.28%

(d) Project age

< 6 mo	6 mo - 1 yr	1 yr - 2 yrs	> 2 yrs
13.33%	18.58%	21.27%	46.82%

(e) VCS used predominantly

Git	SVN	Hg	Microsoft TFS	CVS	Other
52.68%	20.37%	12.07%	8.54%	1.10%	5.24%

Classification of open-ended questions.

The survey contained both multiple choice and open-ended questions[1]. We hand-coded the answers to the open-ended questions using *qualitative thematic coding* [18]. We developed a set of codes that we validated by achieving an inter-rater agreement of over 80% for 20% of the data. Two coders, the first and the third authors, developed the categories which were not known apriori. For measuring the agreement we used the Jaccard coefficient.

2.2 Repository

To provide further insights into how DVCS affects developer's practices, we collected and analyzed 132 software repositories.

2.2.1 Repository Corpus

To answer our research questions we needed to collect repositories that are representative of the centralized and distributed paradigms. We also collected hybrid repositories that started in a centralized paradigm and switched to the

[1]Survey questions can be accessed on our website at `http://cope.eecs.oregonstate.edu/VCStudy/`

Table 2: Repository corpus.

Repo. Type	Repositories	Commits	Authors	Total LOC changed
SVN	52	95571	451	270M
Hybrid	29	151004	2249	89M
Git	51	111725	3190	50M
Total	132	358300	5890	409M

distributed one. Our assumption is that differences in metrics taken from these 3 kinds of repositories provide valuable insights on how they influence source code management.

Using the survey results, we selected SVN and Git as being representative for the centralized and distributed categories, respectively. We collected SVN repositories from SourceForge as and Git repositories from GitHub. These repositories span several programming languages: Java, C, C++, JavaScript, and Python.

For GitHub we selected the top ranked repositories, i.e., repositories that have been marked as favorites by developers and/or have been forked the most. For SourceForge we used its own internal ranking metric to select the top ranked repositories. We queried the SourceForge projects through the Notre Dame Sourceforge Research Archive [8], which serves as a mirror designed specifically for researchers. By choosing the top repositories we ensure that we collect mature projects with a rich history.

To select hybrid repositories, we searched for internet posts about migrating repositories from SVN to Git. In addition, while collecting Git repositories, some of them proved to have actually started in SVN. Thus we classified them as hybrid. We distinguish the two stages of hybrid repositories as Hybrid$_{SVNStage}$ and Hybrid$_{GitStage}$.

We took extra care to ensure the integrity of repositories, i.e., Git repositories did not originate in SVN, by searching for keywords in commit messages.

Table 2 shows the corpus of repositories. For each repository kind, we tabulate the number of individual repositories, commits, and authors that contributed. The last column shows the total number of lines of code that have been changed by all commits.

We aimed for an equal number of SVN and Git repositories to ensure that we compare the two paradigms in a fair way.

2.2.2 *Repository Analysis*

We have built an analysis platform to gather several commit metrics. We used Git as the canonical representation for all repositories. This is possible since the Git object storage model is a superset of the centralized model. For example, the linear history of CVCS can be easily represented in Git's directed acyclic graph branching model. Thus, we converted all SVN repositories to Git, using the svn2git tool [10].

Our platform builds on top of gitective [3], a framework capable of traversing history trees, one commit at a time.

To explore the statistical significance of various sample differences, we applied the Wilcoxon rank-sum test. We chose this test since none of the data fit a normal distribution.

We used the Pearson correlation coefficient in order to establish linear dependence between two sets of randomly distributed values.

Filtering changes.

In our initial manual investigation of commits we have discovered that many commits do not represent actual programming changes carried out by a developer (e.g., adding features, bug fixing, refactoring, etc.), but are the result of applying tools such as code formatters. Such commits are extremely large, i.e., they affect thousands of LOC. Since these commits would bias our analysis, we decided to filter them out. Our analysis filters out any commit that:

- consists only of either added, deleted or renamed files. Most of the times these commits represent large scale project file structure modifications. Also, commits that only add files do not interact with any part of the program and were therefore eliminated.
- are merge commits. These commits usually represent decisions on conflict resolution and contain changes from several lines of development.
- updates only copyright notes, code documentation (e.g., JavaDoc comments) or reorganize code dependences (e.g., `import` statements).
- is artificially created by repository migration tools.

Inside each commit, we ignore all changes that modify only comments and white spaces.

Commit metrics.

For each commit we collect the following metrics.

Commit id, for identifying commits.

Commit date, for sorting commits chronologically.

The author of the commit, for grouping by authors.

Number of LOC changed by the commit, for determining the size of commits. For each commit we compute LOC added, deleted, or modified as reported by the standard diff tool.

Number of files impacted by the commit, for determining the commit size. While LOC tells us how much software editing has been performed in a commit, the number of impacted files tells us how spread the change is within the system.

Number of issues referenced in the commit message, to determine the cohesiveness of changes. The issues refer to programming tasks, such as features or bugs, managed with external systems such as BugZilla, Jira [27], etc. In order to detect them, we used an approach similar to the one described by Bird at al. [15], which employs searching for specific text patterns in the commit message.

3. RESULTS

3.1 How does the VCS type affect developers' behavior?

RQ 1: Does the type of VCS affect the size of commits?

Table 3 shows the commit size, both in lines of code and in number of files, made by individual authors. This data is grouped by VCS type.

In terms of LOC, the commits from Git repositories tend to be smaller than those made in SVN repositories ($p < 0.01$). The mean and median lines of code changed per commit in Git repositories is 27.20 and 13.46, respectively, while for SVN repositories these values are 40 and 18.44 respectively. The standard deviation of changed lines of code also

Table 3: Commit size across different VCS

	Mean		Median		StdDev	
	LOC	files	LOC	files	LOC	files
Git	27.20	3.08	13.46	1.96	32.72	2.7
Svn	40.06	5.65	18.44	3.19	49.62	6.72
HybridGit	23.02	2.40	11.52	1.70	27.57	1.74
HybridSVN	25.72	2.82	12.61	1.96	31.24	2.15

differs, with 32.72 lines of code for Git and 49.62 lines of code for SVN.

In terms of the number of files that are affected by a commit, the same trend of a smaller commit size can be seen as with the commit lines of code, although the difference is not significant (p = 0.2). Commits from Git repositories tend to affect fewer files than commits from SVN repositories.

On the other hand, hybrid repositories do not show a smaller commit size after they transition to Git (p>0.5).

Observation 1: DVCS repositories have a smaller commit size than CVCS repositories, in terms of lines of code.

Observation 2: Hybrid repositories do not show any difference between the size of commits performed before and after the switch to the distributed paradigm.

Interpretation: One possible explanation for Git commits being smaller than SVN commits is the fact that Git enables its users to select finer grained changes to commit. In Git the atomic unit of change that can be committed is the line while in SVN it is the file.

Another possible cause that enables small commits in Git is that each developer commits to his own local repository without the need to synchronize with everybody else. This means that there is no risk of conflicts upon every commit.

One participant stated that *"Git promotes the idea that your commit space is not inflicting pain on anyone else, so frequent commit and experimentation is encouraged. By design it promotes small, frequent commits that serve a specific purpose rather than the '5pm commit.'"* Resolving conflicts becomes a task that is consciously entered into when deciding to synchronize changes with other team members. It is not something that must happen with every commit.

Hybrid repositories on the other hand do not seem to experience smaller commits after switching to Git, as observation 2 shows. Our assumption is that in such cases a certain commit policy is formed within the team while the project is under SVN. This commit policy is then intuitively carried over after switching to Git, leading to the same observed commit size.

The culture of the project takes a longer time to change when a new tool is introduced. Thus, in long lasting projects, it seems that old habits die hard.

RQ 2: Do developers split their commits into logical units of change? How do they do it?

The changes that a developer makes might belong to one or more logical units of change. Do developers split these changes and commit separately? Or do they just group everything and generate one large commit? The answers in the survey give us the picture depicted in Table 4.

Table 4: Developers splitting their commits (%)

Practice	DVCS	CVCS	Overall
Split their changes	81.25	67.89	75.99
Group their changes	12.50	26.61	18.05
Other	6.25	5.50	5.96

Table 5: Reasons for splitting commits (%)

Technique	DVCS	CVCS	Overall
Implementation details	37.01	21.85	32.03
Intent of change	45.13	62.251	50.76
Policy	6.17	5.30	5.88
Other	11.69	10.60	11.33

Observation 3: 76% of the developers split their commits. The percentage is higher for distributed version control systems (81.25%), compared to centralized ones (67.89%).

One explanation for this fact is that in DVCS, the commit process is easier and cheaper than in centralized ones. There is no risk of conflict with each new local commit. Moreover, the smallest atomic unit of change in DVCS is the line, not the file (as it is in CVCS). All these make committing easier, so developers are willing to take the time to split and commit each logical change separately. In a recent study conducted in parallel with ours, Muslu et al. [32] have also discovered that the ability to commit locally and independently allows developers to work incrementally.

A question of great interest is the criteria on how they split their changes. We chose four categories to capture the respondents' answers:

Implementation details refer to *how* was a change carried out (e.g., change field type, add new branch to a switch statement, etc). *Intent of change* splits changes by expressing the *what* part of the change carried out (e.g., add a feature, fix a bug). *Policy* splits changes based on a criteria that is externally imposed (management practices, development process, etc). *Other* represent reasons that do not fit in the above criteria.

Table 5 tabulates the reasons for splitting commits.

We observe that in the case of DVCS, developers split their changes based on implementation details more frequently than they do in CVCS. This will inevitably result in more commits. As is the case with observation 3, we can attribute this to an easier commit process.

Observation 4: Overall, developers choose to split their commits using the intent of change.

Observation 5: More DVCS users split changes based on implementation details than CVCS users.

As we have seen in observations 3 and 5, DVCS users split their changes in several commits more often and they do it with a finer-grained scope in mind. One participant reported: *"Each commit is one cohesive change that might fix a bug, add new functionality, alter existing functionality ([...] like "sphere class can now calculate its own volume" - user level features usually take many commits)"*. This corroborates with the findings about the influence of Version

Control Systems on commit size (RQ 1). Being able to more easily split the commits and the commit process being simpler as well, will result in smaller commit size.

RQ 3: How often and why do developers squash their commits?

Squashing refers to the operation of merging two or more commits into a single one.

Results from the survey show that only 30% of the developers squash their commits. The results for the distributed and centralized repositories are shown in Table 6.

Table 6: Developers squashing their commits (%)

Response	DVCS	CVCS	Overall
Yes	36.59	18.12	30.21
No	54.79	44.57	51.31
Not applicable	8.62	37.32	18.48

Table 7: Reasons why developers squash their commits (%)

Reason	DVCS	CVCS	Overall
Group similar changes	25.63	45.16	28.80
Intermediate steps are irrelevant	20	0	16.75
Remove mistakes	15	0	12.57
Keep history clean	26.88	6.45	23.56
Policy requirement	5.63	9.68	6.28
Other	6.88	38.71	12.04

Table 6 shows that squashing happens twice more often in distributed repositories than in centralized ones[2]. This probably has to do with the fact that it is easier to manipulate commits in DVCS. Developers who practice squashing mention two main reasons (Table 7): (i) to group several changes together and; (ii) they do not care about the path they took to a solution as long as it's finished and it works.

Observation 6: Squashing does not occur often in practice. If it does occur, it's a practice mainly associated with DVCS

RQ 4: Why do developers prefer one Version Control System over another?

According to the survey, we have found two main reasons why developers find a commit process more natural. The first is the presence of a *killer feature*. It usually helps developers achieve higher productivity by allowing a workflow that is more comfortable for them. The second is habit. Developers get used to a certain tool. Therefore, they will find the tool natural to use from the habits they have acquired while using it on a daily basis. Table 8 summarizes the complete results.

In 46% of the cases developers prefer DVCS because of a *killer feature*. By looking at individual replies we have found that one of the features mentioned is the possibility to commit to the local copy of the repository. Also, we can see that the main reason for preferring CVCS is the ease of

[2]See Internal Threats to Validity (Section 4)

Table 8: Reasons for considering a VCS more "natural" to commit (%)

Reason	DVCS	CVCS	Overall
Killer feature	46.02	10.89	30.41
Old habit	22.88	41.58	30.41
Easy to use	19.79	41.58	27.14
Personal preference	2.06	0.99	2.04
Other	9.25	4.95	10

use. While the distributed model has its advantages, that comes at the cost of a more complex model. This could explain why so many developers (almost 42%) think that the centralized model is easier to use.

Also, many prefer CVCS simply because of habit. Having used a system for a very long time, one gets used with the command interface and paradigm. It is interesting to note that CVCS are used not for their capabilities in managing change, but for old habits and a faster learning curve.

Observation 7: The commit process of DVCS is perceived by developers to be more natural because of the presence of killer features.

Observation 8: The commit process of CVCS is perceived to be more natural because of familiarity and a faster learning curve, not their feature set.

Our findings are reinforced by Muslu et al. [32]. Their study shows that developers prefer the ability to work offline. Also, they have found the learning curve to be a barrier in adopting DVCS.

RQ 5: Does the VCS influence the frequency with which developers commit?

Table 9 shows results we obtained from the survey. Developers commit several times a day regardless of the version control they use. The data for each VCS type shows a slightly different picture. Developers using DVCS commit once an hour more often (19.66%) than developers using CVCS (4.10%). Also, when using CVCS developers are more likely to commit once a day (14.75%) than when using DVCS (7.19%).

Observation 9: Most developers have similar habits independent of what VCS they use.

Table 9: How often do developers commit? (%)

	DVCS	CVCS	Overall
Once a minute	3.38	0.82	2.51
Once an hour	19.66	4.10	14.37
Several times a day	65.96	66.80	66.25
Once a day	7.19	14.75	9.76
Several times a week	1.90	9.43	4.46
Once a week	1.48	3.32	2.09
Once a month	0.42	0.82	0.56

Interpretation: The fact that developers commit once an hour more often when using DVCS than when using CVCS suggests that they find it easier to commit. Results from the previous research questions also lead to this conclusion. One interesting results is that 14.75% of developers using CVCS commit once a day. This suggest a pattern of committing once the work day is over.

Implications.

For developers: Smaller commits make code reviews easier. Having a tool that enables small, fine grained commits allows users to separate and document each change individually. One participant mentioned that they split their commits because *"[changes] should be logically separated, to easily allow [the] commit message to drive [the] review"*. Consider reviewing a new feature that has been added. Instead of going through thousands of changes, the reviewer can go through one change at a time, each explained by the commit message.

Also, smaller commits enables easier bisecting. This enables techniques such as Delta Debugging [38] to be employed to find the root cause of bugs.

Using a DVCS can offer developers more power when it comes to choosing what to commit. DVCS tools like Git allow the splitting of commits at line level, which helps when changes with multiple intents are interleaved in a single file. This kind of separation is not possible when using SVN. A participant mentioned that he preferred Git because *"it gives useful tools for splitting or merging commits"*.

By splitting changes into multiple and smaller commits developers can cherry-pick changes. Cherry-picking refers to the operation of selecting one commit from a branch and applying it to another one. This way, developers can migrate changes from one branch to the other without the need to merge all changes. This has maximum benefits when commits carry only one intent, as noted by one respondent who splits his commits because of *"the ability to easily cherry pick or revert [commits]"*.

Developers can remove mistakes and clean a project's history by squashing their commits. Several respondents mentioned that they squash to *"To correct a previous commit"* or *"To make it easier for people reading the log to understand what's been changed"*. However, we see in observation 6 that it is not widely used. This is because, sometimes, squashing leads to a loss of historical data. This information might be useful in the future when debugging or trying to understand the origin of some changes.

From observation 7 we learn that developers like DVCS because of some of their *killer features*. One that was mentioned often was the ability to commit locally: *"You get to commit to a local repository and make your changes public only when they are ready"*. Learning how to use these features takes time and effort. Using the same tools allows developers to keep their level of productivity in the short run. However, the initial effort and loss of productivity caused by learning a new version control system or paradigm may pay off in the long run. One participant reported that he *"tried Git but its too similar yet just different enough to confuse the hell out of me and slow us all down"*. Another *"[...] was not happy about this [using Git] to start off with, and it took me about two years to learn and love Git"*. The advantages of switching would be overall increased productivity, compared to using a CVCS, and better history and management of software changes.

For researchers: Researchers mining software repositories and studying discrete changes should focus on DVCS because they allow smaller atomic units of change. For refactoring researchers, the smaller Git commits could better define individual changes. For researchers who tie different software artifacts to code, such as bugs, Git commits are more precise therefore they may have fewer false mappings.

Researchers must be careful when collecting software repository related metrics. We have found that old repositories that migrated through several VCS tools present a different behavior than pure repositories. It may be the case that the culture formed in the era of the first VCS shadows the subsequent ones. There might other phenomena that influence a repository's structure. By not paying attention to different phenomena that affect repositories researchers risk biasing or confounding their results.

There is a lot of noise when studying different types of software changes introduced by commits. As seen in section 2.2.2, there are many types of commits and individual changes that do not constitute acts of development. Researchers should clearly define what types of changes they are studying and then take the appropriate actions to filter undesired commits. By not paying attention to different types of commits, researchers risk biasing or confounding their results.

DVCS allow users to change history before they make it public or available to others. One participant stated he squashed commits because he *"committed more often locally while working. That need not be seen in the final push, because it usually only adds noise"*. This is a threat when mining repositories. The repository that is publicly available might not be the one that developers had when they committed their changes. Squashing is just one of the ways in which developers can change history. Research on such repositories should take this threat into account.

For tool builders: Although Git enables finer grained changes, it is still the developers' task to disentangle these changes. This is a manual, tedious, and time consuming process. VCS tools could keep track of different change intents and then offer to commit them separately. Herzig et al. [24] show a technique by which this can be achieved. They devised a heuristic untangling algorithm that splits tangled changes according to different source code criteria (e.g., the distance between two changed AST nodes).

We envision a new generation of tools that can use the average size of a commit as a quality metric. When a developer has uncommitted code larger than a threshold, the tool could suggest that it's time to split changes and commit.

Continuing on the idea of metrics, the field of software design flaws can be applied to repositories as well. Researchers have identified many software design flaws [28]. Marinescu [29] presents detection strategies for these flaws, allowing tools to identify, report and offer suggestions for improvement. By following this approach researchers can devise design flaws for repositories and then metric based detection strategies for these flaws would allow tools to measure the health of a repository.

Squashing is a process by which history can be altered or completely lost. To prevent history loss, VCS tools could support features such as hierarchical commits: the ability to create a virtual commit that holds other real commits. Instead of loosing history through squashing, developers could group commits into larger, composite commits.

Our finding that developers from hybrid repositories use the same habits after switching to DVCS as when they used

CVCS suggests the need for tools to help educate developers on how to effectively change their habits.

Respondents identified features as an important factor for using DVCS. Some mentioned that certain features were an integral part of their workflow. Paying attention to these workflows and creating the tools to support them will pay off in the future. The payoff will increase productivity on the developers side, and bring a larger user base on the tool builder's side, since developers will prefer a tool that best fits their work style.

For team management: As Observation 2 states, hybrid repositories do not show the same trends as non hybrid ones. Adopting new tools and new technology is only part of the change and by no means enough or complete. Tools that bring a new vision to how software is developed should be followed by a shift in policy and project culture as well. One cannot hope to improve the development process by only improving the tools.

3.2 How does the team size affect VCS usage?

RQ 6: Does team size affect the choice of VCS?

Table 10: VCS choice by team size (%)

VCS type	1	2-5	6 - 10	11 - 25	26 - 100	> 100
DVCS	82.22	72.07	62.30	65.22	60.00	70.97
CVCS	17.78	27.93	37.70	34.78	40.00	29.03

Table 10 shows that most teams use the distributed model, regardless of the team size. However, the presence of the centralized version control systems increases once the team size increases.

Observation 10: Teams of all sizes predominantly prefer DVCS

Interpretation: Since 53% of the survey projects are less than two years old, it is likely that they were developed during the rising popularity of the DVCS. As for the popularity of CVCS at larger teams, this can be interpreted as inertia of larger teams to use new paradigms and tools. However, once the team size crosses 100, the overhead of merging changes pushes the team against their inertia.

RQ 7: Are larger teams more likely to use Issue Tracking Systems (ITS)?

We are answering this question using two data sources, the survey and the repository analysis.

Table 11 summarizes results collected from the survey.

Overall, 93.87% of the participants reported using an ITS. While the value is lower for one-person projects (63%), it is over 90% in all other cases. More interestingly, all participants working in projects with over 100 developers report using an ITS.

Observation 11: Most projects use an Issue Tracking System.

The analysis of the repositories shows that 91.6% of the projects contain commit messages that refer to issues in an ITS. This correlates with the survey.

Table 11: Issue Tracking System (ITS) usage based on team size (%)

Team size	Use an ITS	Don't use an ITS
1	63.04	36.96
2 - 5	97.88	2.12
6 - 10	92.23	7.77
10 - 25	95.24	4.76
26 - 100	97.01	2.99
Over 100	100.00	0.00
Overall	93.87	6.13

RQ 8: Does team size affect the size of commits?

We measured the correlation coefficient between team size and commit size (measured in lines of code and number of changed files). Table 12 shows no linear relation between team size and the size of commits. This holds for both LOC and number of files. The fact that all but one coefficient are negative could indicate a weak downhill trend for commit size as team size increases.

Table 12: Correlation coefficients between team size and commit size

Repository type	LOC	# of files
Git	-0.11	-0.09
SVN	-0.07	-0.16
Hybrid$_{GitStage}$	-0.01	0.16
Hybrid$_{SVNStage}$	-0.06	-0.35

Observation 12: There is no discernible relationship between the team size and the size of commits other than a weak tendency for commit size to decrease as team size increases.

Interpretation: We expected to see that commit size decreases as team size increases. Small teams can be very agile and quickly grow the code base because everybody knows what everybody else is doing. Large teams have less knowledge of the overall task distribution, and they must exercise more care when accepting and integrating changes from many sources. Therefore, we assumed, large teams would perform smaller commits to better express changes.

However, the data shows that developers use the same intuition for splitting changes, regardless of team size. This could hint that large teams must use other mechanisms to control the complexity of software changes.

Such mechanisms could be more complex branching and merging models. Philips et al. [33] show that as projects grow in size and activity, the complexity of the branching model increases.

RQ 9: Does team size influence commit squashing?

Another question is whether developers working in larger teams squash more frequently than developers working in smaller teams. Table 13 shows that this could be the case. While teams of two to five developers squash only 27% of the time, teams with over 100 developers squash 57% of the time. This is more than double the rate compared to smaller teams. The reasons for this are twofold:

Table 13: Squashing in relation to the team size (%)

Team size	Squash	Don't squash	Not Applicable
1	17.39	56.52	26.09
2-5	27.27	54.55	18.18
6-10	42.25	29.50	28.06
11-25	30.16	50.79	19.05
26-100	40.00	46.15	13.85
Over 100	57.14	34.29	8.57

1. According to RQ 6, larger teams tend to use CVCS more often.
2. Developers might try to keep the upstream history clean. Working in a large team means that if every developer were to push their full history, the main repository could get cluttered. So squashing would mitigate an explosion in the number of commits and branches.

Observation 13: Large teams squash commits more often.

Implications.

For researchers: Best quality data about issue tracking systems is obtained from projects developed by large teams.

On the other hand, large teams tend to squash more often which would result in bigger commits with more entangled changes. Therefore, researchers may have to trade off some traits over other ones when choosing to study repositories from small or large teams.

For tool builders: Practically all large teams use Issue Tracking Systems in order to track work items. However, in the current state of practice developers track code and issues by inserting issue references inside commit messages. This is tedious and imprecise since developers have to manually group changes by issue.

Therefore, tool builders should create tools that (i) keep track of code written for a particular task, (ii) automatically group code changes by issue, and (iii) incorporate issue tracking inside Version Control Systems.

Tool builders should abstract away low level VCS concepts in order to ease the learning curve of Version Control Systems. Instead of using implementation level details (branch, rebase, hash, etc) as their interface to the user, VCSes could use a high level vocabulary from the domain of software processes (feature, bug, review, etc.).

For team management: Very large teams that struggle with aggregating changes should consider investing the extra effort and switch to distributed tools (Git, Hg, etc) and more involved branching models (specialized branches, deeper forking tree, etc) [33].

3.3 How does the VCS type affect the development process?

RQ 10: Does the type of VCS influence the presence and the number of issue tracking labels (ITL)?

The survey shows that the majority of developers (69%) commit only one issue at a time (Table 14). This figure is slightly smaller for CVCS than it is for DVCS. However, CVCS developers perform commits with more than one issue more often (17%) that developer using DVCS (9%).

Table 14: Survey: How do developers commit if they work on more than one issue (%)

	1 issue	>1 issue	Not applicable
Distributed	68.71	8.59	22.7
Centralized	66.67	17.03	16.03
Overall	69.25	11.13	19.13

The repository analysis shows that 31% of all commits contains an ITL. This mirrors similar findings by Bird et al. [15]. The number is higher for Git repositories at 43.42%. For hybrid repositories the number is at 33.12% compared to SVN at 13.13%.

Observation 14: A small number of commits are labelled with ITL. Nevertheless, issue tracking labels appear more frequently in DVCS commit messages than in CVCS commit messages.

Interpretation: The fact that we see a larger number of developers committing changes belonging to two or more issues per commit for CVCS might be an indication of a higher difficulty in selecting the changes to be committed. The difference in the granularity of change selection between the tools could explain these results.

The overall low number of commits with ITL can be attributed to a relaxed commit policy. As the repositories are gathered from open source projects, it may be difficult to enforce a strict commit policy. To the best of our knowledge, none of the analyzed repositories enforced a practice of mandatory ITL inclusion in commit messages.

RQ 11: Is there a correlation between the number of issue tracking labels in the commit message and the commit size?

We can see that commits to SVN and Hybrid repositories tend to be larger when more issues appear in the commit message. The correlation is strong and positive, at 0.68 for SVN and 0.81 for Hybrid repositories. For Git repositories this trend does not hold. There is a slight tendency for commit size to decrease when the number of issue tracking labels increases. There is a weak negative correlation, at -0.38). Table 15 shows the detailed results.

Table 15: Average LOC for issue references by VCS type

VCS type	number of issue references					corr.
	1	2	3	4	5	
SVN	33.04	35.69	54.56	31	80	0.68
Git	25.27	36.46	38.05	23	23	-0.38
Hybrid	27.67	31.66	37.74	83.2	62.14	0.81
All	28.59	34.72	39.91	57.78	60.08	0.97

Observation 15: In SVN and Hybrid repositories commit size is positively correlated with the number of referenced issues.

Observation 16: In Git repositories there is a weak negative correlation between the commit size and the number of referenced issues.

Interpretation: The strong correlations for SVN and hybrid repositories reinforce the idea that in these repositories developers tend to group different change intents (issues) together.

In Git, the trend seems to be opposite. This suggests that Git commits do not get larger in size when they reference several ITL. Rather, Git commits could contain a change common to all referenced issues.

Observations 3 and 5 show that developers using DVCS split their commits more often and that their commits contain finer scoped changes. This hints to the idea that they might carve out the common piece of code that contributes to solving both issues.

RQ 12: How does the size of commits vary in time?

In order to investigate how the commit size varies in time, we averaged the commit size for monthly intervals. We then calculated correlation coefficients for the monthly values of average commit size.

Table 16 shows that the commit size tends to become smaller as projects get older. The average age of a typical repository (time between the first and the last commit) is 55 months. For SVN repositories it is 54 months, for Git repositories it is 30 months and for hybrid ones it is 94 months.

Table 16: Correlation between commit size and commit time

VCS	average corr.	# of positive corr.	# of negative corr.	% of negative corr.
SVN	-0.06	21	31	60%
GIT	-0.17	13	25	66%
Hybrid	-0.11	12	16	57%
ALL	-0.11	46	72	61%

The average commit size usually decreases by approximately 15-20% during the lifetime of a repository. Overall correlation between commit size and time of commit for all types of repository is usually negligible (-0.11) and in most cases appears to be negative.

Commit size tends to decrease more in Git then it does in SVN. 71% of the analysed Git repositories show decreasing trends in average commit size over time. For SVN only 60% showed this trend, while for hybrid repositories this number constitutes only 57%.

Observation 17: Commit size tends to become smaller as projects get older.

Interpretation: This decreasing trend can be explained by different types of changes that happen during projects' life. In the early stages of development, commits tend to be larger because developers are adding features from scratch. As the project matures, development switches from adding new features to performing corrective changes, like bug fixes. Corrective changes are usually smaller in size.

Implications.

For researchers: We found that the average commit size slightly decreases over time. This could be explained by a variation of change practices during software development stages (development, testing, support, etc). A more detailed investigation on how software development stages influence change practices is needed.

DVCS commits contain more ITL than CVCS. This suggests that DVCS repositories are better candidates for research projects studying links between commits and issue tracking systems.

For tool builders: As long as changes tend to become smaller as projects mature, it becomes worthwhile to rebuild smaller parts of applications as changes occur. Thus, we encourage tool builders to provide more support for intra file incremental builds.

Changes are more granular in DVCSes and usually have only one issue reference. VCS tool builders could include new abstractions that represent features. For example, cherry picking could be done at feature level.

One of the reasons why developers do not put issue numbers into commits could be the extra work it involves. Building upon the suggestion of section 13 for tool builders, there could be a better integration between VCS and issue tracking systems.

For team management: Because commit size tends to become smaller as projects get older, it is reasonable to assume that developers tend to spend more time analyzing existing source code instead of adding new code. Therefore developers' productivity should be measured not only by the amount of code, but also by the complexity and importance of their changes.

4. THREATS TO VALIDITY

Construct: Are we asking the right questions? We are interested in assessing the state of the practice for version control systems usage. Thus, we think that our research questions have high potential to provide a unique insight and value for different stakeholders: developers, researchers, tool builders and team management.

Internal: Is there something inherent to how we collect and analyze the VCS usage that could skew the accuracy of our results?

One of the main threats is the practice of squashing commits. As we have shown in RQ 3, squashing is a used practice among software developers. For DVCS, roughly 36% of developers squash their commits. Because squashing rewrites history, it is impossible to detect squashing activity. The main effect is that commits gets larger, because squashing combines two or more commits into a single commit. The result is an increased commit size. Thus, the average commit size for DVCS might be even smaller than the ones we report. Observation 1 would still stand even in the case of heavy squashing practices.

Another threat is that our results may be biased by the development culture. As Rigby et al. [35] mention, commits done to Open Source Software (OSS) tend to be smaller than the ones done in proprietary software. However, both our SVN and Git repositories are originating from the open-source community, so the OSS culture would affect both in similar ways.

Also, there is a chance that one developer might use different name aliases when committing in the same repository.

This could affect metrics that rely on team size for repositories analysis. Even though we have encountered few cases of aliases usage while analyzing repository data, we also noticed that it is a very rare and exceptional practice.

While designing our survey we aimed at keeping it short. However, in doing so, some of the participants may have misunderstood our questions. For example, when we asked the question *"Do you squash your commits?"* we were aiming to find if developers are using the *squash* command from Git or similar tools. This command collapses together commits *after* they were committed. However, respondents might have interpreted the question as squashing multiple changes *before* committing. This can explain why 18% of developers using CVCS reported that they use squashing, even though this is not possible in CVCS tools. While we did run a pilot [18] of our survey, there is always the possibility that we have miscommunicated our intent.

One other threat is the possibility of age bias in our repositories. Since SVN has been available for a longer period of time, SVN repositories might contain older, more mature projects than Git repositories.

External: Are our results generalizable for the general version control usage practice?

While we analyzed 132 repositories from the open source community, we cannot guarantee that these results will be the same for proprietary (closed source) software. However, given the overall agreement between the survey, which was filled in mostly by developers working with proprietary software, and the data we acquired by analyzing the repositories, we can assume that the general trends that we found will be true for proprietary software as well.

In our corpus of open-source repositories, 83% of the projects were developed in Java, and the remaining 16% used C/C++, Javascript and Python. Moreover, the pure Git repositories consist of 98% Java projects whereas the pure SVN repositories consist of 80% Java projects. While we have no reasons to believe that programming language affects the culture of committing changes, in the future we plan to diversify our corpus and explore change variation in programming languages.

The sources for our repositories are GitHub and SourceForge. This means that we only looked at projects that used Git or SVN. We did not study other VCS tools for the distributed or the centralized paradigm. However, as the data from our survey indicates, Git and SVN are the predominant systems used today. They are the most widely used in their class, thus we think they are representative.

Reliability: Can others replicate our results? The list of repositories we used for our analysis is available online [11]. Also, the infrastructure we used for the analysis is available open source as a GitHub repository [5].

5. RELATED WORK

To the best of our knowledge, our paper is the first study to *compare* the impact of CVCS and DVCS on the practice of committing changes.

Several researchers [12, 13, 20, 23, 25, 26, 30, 34, 34] studied the practice of commits but only in the CVCS paradigm. Purushothaman et al. [34] and German et al. [20] and Hindle et al. [25] studied the properties of typical small commits or typical large commits. Hattori et al. [23] study the size of commits with the purpose of classifying changes. Arafat et al. [13] studied the distribution of commit size. Hofmann et al. [26] predict commit size based on commit history. Herzig et al. [24] propose an algorithm for untangling changes in CVCS. However, ours is the first study to compare the commit size in CVCS and DVCS.

Related with our study about the impact and the presence of issue tracking systems (ITS), several researchers [14, 16, 22, 31, 37] studied ITS. Tian et al. [37] and Bird et al. [16] aim to link source code with ITS. Meneely et al. [31] makes suggestions on improving issue tracking labeling in commit messages. Bachmann et al. [14] mine bug tracking databases with the purpose of linking ITS with the software development process. Hassan et al. [22] provide an overview of repository and ITS mining practices. However, none of these studies compared CVCS and DVCS based on ITS practices.

Recently, researchers started mining DVCS repositories for collaboration between developers [36], processes [17, 21] for mining DVCS repositories

6. CONCLUSIONS

In this paper we present the first in-depth study to measure the impact of DVCS on software change. To this end we ran a survey with 820 participants and analyzed a corpus of 132 repositories.

We found that the use of CVCS and DVCS have observable effects on developers, teams and processes. The most surprising findings are that (i) the size of commits in DVCS was smaller than in CVCS, (ii) developers split commits (group changes by intent) more often in DVCS, and (iii) DVCS commits are more likely to reference issue tracking labels. These show that DVCS contain higher quality commits compared to CVCS due to their smaller size, cohesive changes and the presence of issue tracking labels.

The survey provided valuable information on why developers prefer one paradigm versus the other. DVCS are preferred because of *killer features*, such as the ability of committing locally. In contrast CVCS are preferred for their ease of use and faster learning curve.

We hope that our work inspires future research not only into the impact that centralized and distributed VCS tools have on software development but also on how general properties of VCS tools enable developers to manage and express change.

7. ACKNOWLEDGEMENTS

We would like to thank Cosmin Rădoi, Alexandru Gyori, David Hartveld, Alex Groce, Michael Rosulek, Will Jernigan, Faezah Bahmani, Iftakar Ahmed, Michael Hilton, Irwin Kwan, Charles Hill, Amber Horvath, Paul McKenney, and the anonymous reviewers for feedback on earlier drafts of this paper.

Also, we would like to thank Joel Spolsky, Robert Martin, and Steve Berczuk for helping us promote the survey.

This research is partly funded through NSF CCF-1213091 and CCF-1219027 grants, a SEIF award from Microsoft, and a gift grant from Intel.

8. REFERENCES

[1] Cvs. `http://cvs.nongnu.org/`. Accessed February 27, 2014.

[2] Git. `http://git-scm.com/`. Accessed February 27, 2014.

[3] gitective: Git repository analysis tool. `https://github.com/kevinsawicki/gitective`. Accessed September 6, 2013.

[4] Github. `http://www.github.com/`. Accessed February 27, 2014.

[5] http:/www.github.com/caiusb/gitsvn.

[6] Mercurial. Accessed February 27, 2014.

[7] Sourceforge. `http://www.sourceforge.net/`. Accessed February 27, 2014.

[8] Sourceforge research data archive (srda): A repository of floss research data. `http://srda.cse.nd.edu/mediawiki/index.php?title=Main_Page`. Accessed September 6, 2013.

[9] Svn. `http://subversion.tigris.org/`. Accessed February 27, 2014.

[10] svn2git: Svn to git repository conversion tool. `https://github.com/nirvdrum/svn2git`. Accessed September 6, 2013.

[11] Vcs usage study companion. `http://tiny.cc/VCStudy`.

[12] A. Alali, H. Kagdi, and J. I. Maletic. What's a typical commit? a characterization of open source software repositories. In *Program Comprehension, 2008. ICPC 2008. The 16th IEEE International Conference on*, pages 182–191. IEEE, 2008.

[13] O. Arafat and D. Riehle. The commit size distribution of open source software. In *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on*, pages 1–8. IEEE, 2009.

[14] A. Bachmann and A. Bernstein. Data retrieval, processing and linking for software process data analysis. *University of Zurich, Technical Report*, 2009.

[15] C. Bird, A. Bachmann, E. Aune, J. Duffy, A. Bernstein, V. Filkov, and P. Devanbu. Fair and balanced?: bias in bug-fix datasets. In *Proceedings of the the 7th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering*, pages 121–130. ACM, 2009.

[16] C. Bird, A. Bachmann, F. Rahman, and A. Bernstein. Linkster: enabling efficient manual inspection and annotation of mined data. In *Proceedings of the eighteenth ACM SIGSOFT international symposium on Foundations of software engineering*, pages 369–370. ACM, 2010.

[17] C. Bird, P. C. Rigby, E. T. Barr, D. J. Hamilton, D. M. German, and P. Devanbu. The promises and perils of mining git. In *Mining Software Repositories, 2009. MSR'09. 6th IEEE International Working Conference on*, pages 1–10. IEEE, 2009.

[18] D. S. Cruzes and T. Dyba. Recommended steps for thematic synthesis in software engineering. In *Empirical Software Engineering and Measurement (ESEM), 2011 International Symposium on*, pages 275–284. IEEE, 2011.

[19] K. Finley. Github has surpassed sourceforge and google code in popularity. `https://www.readwriteweb.com/hack/2011/06/github-has-passed-sourceforge.php`, 2011. Accessed February 27, 2014.

[20] D. M. German. An empirical study of fine-grained software modifications. *Empirical Software Engineering*, 11(3):369–393, 2006.

[21] G. Gousios and D. Spinellis. Ghtorrent: Github's data from a firehose. In *Mining Software Repositories (MSR), 2012 9th IEEE Working Conference on*, pages 12–21. IEEE, 2012.

[22] A. E. Hassan. The road ahead for mining software repositories. In *Frontiers of Software Maintenance, 2008. FoSM 2008.*, pages 48–57. IEEE, 2008.

[23] L. P. Hattori and M. Lanza. On the nature of commits. In *Automated Software Engineering-Workshops, 2008. ASE Workshops 2008. 23rd IEEE/ACM International Conference on*, pages 63–71, 2008.

[24] K. Herzig and A. Zeller. The impact of tangled code changes. In *Proceedings of the Tenth International Workshop on Mining Software Repositories*, pages 121–130. IEEE Press, 2013.

[25] A. Hindle, D. M. German, and R. Holt. What do large commits tell us?: a taxonomical study of large commits. In *Proceedings of the 2008 international working conference on Mining software repositories*, pages 99–108. ACM, 2008.

[26] P. Hofmann and D. Riehle. Estimating commit sizes efficiently. In *Open Source Ecosystems: Diverse Communities Interacting*, pages 105–115. Springer, 2009.

[27] J. Janák. Issue tracking systems. *Brno, spring*, 2009.

[28] M. Lanza and R. Marinescu. *Object-oriented metrics in practice: using software metrics to characterize, evaluate, and improve the design of object-oriented systems*. Springer-Verlag New York Inc, 2006.

[29] R. Marinescu. Detection strategies: Metrics-based rules for detecting design flaws. In *Software Maintenance, 2004. Proceedings. 20th IEEE International Conference on*, pages 350–359. IEEE, 2004.

[30] M. Marzban, Z. Khoshmanesh, and A. Sami. Cohesion between size of commit and type of commit. In *Computer Science and Convergence*, pages 231–239. Springer, 2012.

[31] A. Meneely, M. Corcoran, and L. Williams. Improving developer activity metrics with issue tracking annotations. In *Proceedings of the 2010 ICSE Workshop on Emerging Trends in Software Metrics*, pages 75–80. ACM, 2010.

[32] K. Muslu, C. Bird, N. Nagappan, and J. Czerwonka. Transition from centralized to distributed vcs: A microsoft case study on reasons, barriers, and outcomes. In *ICSE '14: Proceedings of the 2012 International Conference on Software Engineering*. IEEE Press, 2014.

[33] S. Phillips, J. Sillito, and R. Walker. Branching and merging: an investigation into current version control practices. In *Proceedings of the 4th International Workshop on Cooperative and Human Aspects of Software Engineering*, CHASE '11, pages 9–15, New York, NY, USA, 2011. ACM.

[34] R. Purushothaman and D. E. Perry. Toward understanding the rhetoric of small source code changes. *Software Engineering, IEEE Transactions on*, 31(6):511–526, 2005.

[35] P. Rigby, B. Cleary, F. Painchaud, M. Storey, and

D. German. Contemporary peer review in action: Lessons from open source development. *Software, IEEE*, 29(6):56 –61, nov.-dec. 2012.

[36] P. C. Rigby, E. T. Barr, C. Bird, P. Devanbu, and D. M. German. What effect does distributed version control have on oss project organization?

[37] Y. Tian, J. Lawall, and D. Lo. Identifying linux bug fixing patches. In *Software Engineering (ICSE), 2012 34th International Conference on*, pages 386–396. IEEE, 2012.

[38] A. Zeller. Yesterday, my program worked. today, it does not. why? In *Software Engineering-ESEC/FSE'99*, pages 253–267. Springer, 1999.