An Empirical Investigation of Command-Line Customization

Michael Schröder
TU Wien
Vienna, Austria
michael.schroeder@tuwien.ac.at

Jürgen Cito
TU Wien
Vienna, Austria
Massachusetts Institute of Technology
Cambridge, U.S.A.
juergen.cito@tuwien.ac.at

ABSTRACT

The interactive command line, also known as the shell, is a prominent mechanism used extensively by a wide range of software professionals (engineers, system administrators, data scientists, etc.). Shell customizations can therefore provide insight into the tasks they repeatedly perform, how well the standard environment supports those tasks, and ways in which the environment could be productively extended or modified. To characterize the patterns and complexities of command-line customization, we mined the collective knowledge of command-line users by analyzing more than 2.2 million shell alias definitions found on GitHub. Shell aliases allow command-line users to customize their environment by defining arbitrarily complex command substitutions. Using inductive coding methods, we found three types of aliases that each enable a number of customization practices: Shortcuts (for *nicknaming commands*, *abbreviating subcommands*, and *bookmarking locations*), Modifications (for *substituting commands*, *overriding defaults*, *colorizing output*, and *elevating privilege*), and Scripts (for *transforming data* and *chaining subcommands*). We conjecture that identifying common customization practices can point to particular usability issues within command-line programs, and that a deeper understanding of these practices can support researchers and tool developers in designing better user experiences. In addition to our analysis, we provide an extensive reproducibility package in the form of a curated dataset together with well-documented computational notebooks enabling further knowledge discovery and a basis for learning approaches to improve command-line workflows.

KEYWORDS

command line, customization practices, collective knowledge, inductive coding

1 INTRODUCTION

A command-line interface, also called a *shell*, is a textual interface that allows users to interact with the underlying operating system by issuing commands. Expert users, such as system administrators, software developers, researchers, and data scientists, routinely use the shell as it affords them flexibility and the ability to compose multiple commands. They perform a variety of tasks on their systems including navigating and interacting with the filesystem (e.g., `ls`, `mv`, `cd`), using version control (e.g., `git`, `hg`), installing packages (e.g., `apt-get`, `npm`), or dealing with infrastructure (e.g., `docker`). Experts can adapt and play with a multitude of commands and arguments, chaining them together to create more complex workflows. All this versatility introduces a common problem in user interfaces of recognition over recall [37], where users have to recall the particularities of syntax and argument combinations, instead of enabling them to use a more recognizable symbol (as in graphical user interfaces).

A way for these experts to introduce recognizability and customize their command-line experience is to attach distinct names to potentially convoluted, but frequently used, command and argument structures, as well as workflows expressed as compositions of commands. This can be achieved by defining shell aliases. An alias substitutes a given name, the *alias*, with a string value that defines an arbitrarily complex command (or chain of commands). The set of aliases users define provides a window into their preferences expressed as part of their personal configuration. Many users publicly share these configurations on social coding platforms such as GitHub, contributing to a collective knowledge of command-line customizations, which can provide insight into the tasks that expert users repeatedly perform and how well the standard environment supports those tasks.

1.1 Contribution

We see our large-scale analysis on command-line user customizations manifested in alias definitions as a unique window of opportunity to study how the standard environment of the command line could be productively extended, modified, and improved. Our work goes hand in hand with existing efforts to innovate on the experience of command lines that employ techniques from research in systems [23, 43], software engineering and programming languages [10, 54, 55], human-computer interaction [15, 53], and artificial intelligence [1, 24, 31]. Particularly, our extensive qualitative and quantitative analysis, in conjunction with our dataset, form the basis for identifying opportunities for improving command-line experience in the following directions: by characterizing customization practices, we gain a categorical understanding underlying the needs and wants of command-line users; based on our analysis, we identify opportunities for innovation and formulate them as implications, accompanied with concrete scenarios and examples; further, our comprehensive dataset enables the foundation of learning approaches, as part of learning-based program synthesis [7, 44], automated repair [34], and recommendation systems [32]; finally, we also see our results and datasets as a basis for usability research that can impact the design of tools and the future of the shell in general.

We summarize the work in this paper as follows:

- We identified nine **Customization Practices**, grouped into three high-level themes: Shortcuts introduce new names. They can be used for *nicknaming commands* (and correcting misspellings in the process), *abbreviating subcommands* like

`git push`, and *bookmarking locations* for quick navigation. Modifications change the semantics of commands. We can use these types of aliases for *substituting commands*, such as replacing `more` with `less`, for *overriding defaults* to customize commands to personal contexts, which often involves *colorizing output*, and also running certain commands as root by *elevating privilege*. Aliases that combine multiple commands are Scripts. They enable many ways of *transforming data* using Unix pipes, and allow for automating repetitive workflows by *chaining subcommands*.

- A **Curated Dataset of Command-Line Customizations**, consisting of over 2.2 million shell aliases collected from GitHub. We view our dataset as a playground for fine-grained discovery that can benefit researchers, tool-builders, and command-line users; for example, researchers can use this knowledge base to discover which commands are frequently used together and how they are combined, while tool-builders can see how their programs are being customized. We also describe the effective mining technique we used to distill this knowledge, which allowed us to capture almost the whole population (94.09 %) of relevant shell configuration files.
- We formulate **Implications for Improving Command-Line Experience** that go beyond single customization practices to address shortcomings and tie them to existing user experience research. Codifying emergent behavior [14] found in our customizations enables *learning repair rules* and *discovering workflows*. We are able to *uncover conceptual design flaws*, where customizations indicate frustrations with underlying command structures, supporting prior research on potential flaws in the conceptual design of certain commands [38]. Based on the prevalence of highly variable command redefinitions, we propose *contextual defaults*, the ability to suggest different command preferences based on user context [51]. Overall, we find that many customizations deal with the tension of *Interactivity vs Scripting*: commands being used to interactively navigate systems, while at the same time being used within scripts for batch-processing.

We now describe usage and syntax of aliases as a vehicle for customization. We further describe our data collection and coding process, followed by a presentation of customization practices. Finally, we discuss implications for usability and review related work in the broader context of this study.

2 BACKGROUND

A shell is a command interpreter allowing the user to interact with an underlying system. The concept of the operating system shell as an independent process executing outside the kernel originated in Multics [40] and was further developed into the original Unix shell `sh` and its various descendants [27, 50]. The POSIX family of standards defines a Shell Command Language [18, 25], whose standard implementation is still the `sh` utility, but there exist a wide variety of popular POSIX-compliant shells like `bash` or `zsh`. These implementations are free to extend the functionality of the shell, but all share a common subset of core commands and programming language constructs. In this paper, we focus on the built-in `alias` command, available on all POSIX shells.

2.1 Usage and Syntax

The `alias` command allows the user to create *alias definitions*, defining command substitutions. When the shell processes the command line, it replaces known alias names with their defined string values. For example,

```
alias ll='ls -l'
```

defines the *alias name* `ll`, that is replaced by the *alias value* `ls -l`. In this case, `ls` is the standard command for listing directory contents, with the argument `-l` specifying a long-form output format. So the alias `ll` (present in many system configurations) is used to specify a default argument to a commonly used command under a different name.

Alias values can be arbitrarily complex strings and can substitute not only simple commands and arguments, but whole chains of commands. The definition

```
alias ducks='du -cksh * | sort -hr | head -n 15'
```

defines the new command `ducks` by chaining together three different command-line tools in order to return the 15 largest files in the current directory.

In general, an alias definition takes the form

```
alias name=value
```

where `value` can optionally be enclosed in single (`'`) or double (`"`) quotes and `name` can be any identifier that is a valid command name.[1] In particular, the alias name can be an existing command, so a re-definition like

```
alias grep='grep --color=always'
```

is possible.

In the remainder of this paper, we will use the more compact notation `a` → `b` to indicate an alias that replaces the name `a` with the value `b`.

2.2 Dotfiles

Aliases can be entered directly on the command line, in which case they are valid until the shell session ends. To make an alias definition permanent, it is common practice to enter it into a file that is read and executed by the shell on startup. The names of these configuration files differ by shell, but common ones are `.bashrc`, `.zshrc`, or `.profile` and their main difference is the order in which they are executed.[2] Often, aliases are also stored in other files referred to by these startup scripts.

These kinds of files—text-based configuration files that store system or application settings—are also known as *dotfiles*, because their filenames usually start with a dot (`.`) so that they are hidden by default on most Unix-based systems. In recent years, people have started sharing their dotfiles on platforms like GitHub.[3] This has the advantage of being able to sync one's configurations across different machines, and also enables exchange and discovery of configurations between users.

[1] Some shells allow for an alternative alias syntax without the equals sign between `name` and `value`. In this paper we only look at POSIX-compliant alias definitions.
[2] https://www.gnu.org/software/bash/manual/html_node/Bash-Startup-Files.html or https://zsh.sourceforge.io/Doc/Release/Files.html
[3] https://dotfiles.github.io

3 DATASET

Our analysis is based on 2,204,199 alias definitions found on GitHub, collected over a period of two-and-a-half weeks from December 20th 2019 to January 8th 2020.

3.1 Data Collection

Alias definitions can appear in any Shell script, but we anticipated that they would predominantly be found in personal configuration files (like `.bashrc` or `.bash_profile`). Unfortunately, this rules out using some prominent existing datasets for our study [33]: The public GitHub archive on BigQuery,[4] while containing over 1.5 TB of source code, only includes "notable projects" (presumably those with a certain number of stars on GitHub) that additionally have an explicit open source license. This leaves out many of the repositories we are interested in, as users sharing configuration scripts for personal use do not usually add a license file and their repositories are generally not "notable". GHTorrent [16], another popular archive of GitHub data, only contains metadata but not file contents.

Therefore, we found it necessary to write our own tooling to directly collect the data from GitHub ourselves. We used the GitHub Code Search API[5] to find files written in Shell language[6] that contain the string `alias`.

Alas, the GitHub Code Search API comes with its own set of limitations:

(1) only files smaller than 384 KB are searchable
(2) forks are not included
(3) requests are rate limited at 30 per minute and there are additional opaque abuse detection mechanisms that impose further restrictions in an unforeseeable manner
(4) the number of results is limited to 1,000 per search request

The first two limitations do not really affect us, as we are interested in smaller files and do not have to consider forks. The rate limiting, while significantly slowing down the retrieval process, is also not a fatal obstacle. The maximum number of returned search results, however, is a critical limitation. To get around it, we wrote a Python tool called `github-searcher`[7] that uses a clever sampling strategy to vastly increase the number of results we are able to retrieve.

The sampling strategy is based on the GitHub API allowing code search queries to be conditioned on file sizes. For example, the query

```
alias language:Shell size:101..200
```

returns up to 1,000 Shell language files containing the string "alias" that have a file size between 101 and 200 bytes (inclusive). Repeating the search with

```
alias language:Shell size:201..300
```

returns up to 1,000 files of a size between 201 and 300 bytes, and so on. Repeatedly searching with the same search term but different non-overlapping file size ranges allows us to significantly increase our sample of the overall population. Another trick further improves on this: the API gives us an option to sort the results by most or least recently indexed; if we run a search using a specific sort order, then we can effectively double the sample size by repeating the same search with the opposite sort order. Thus we can get up to 2,000 results per search per file size range.

Additionally, while GitHub does not allow us to retrieve more than a limited number of files per query, it does return the total count of files matching the query. While this count is usually very erratic on broad searches, fluctuating wildly between repeated requests, it turns out to be fairly accurate for searches with a small number of results, such as those conditioned on a narrow range of file sizes. This allows us to get a good estimate of the population, and how accurately our sample approximates it.

For this study, using the search term

```
alias language:Shell
```

and the sampling strategy described above, we started by sampling all files in increments of 100 bytes and stopped when we reached 29 KB, about ten times the median file size of the estimated population encountered so far. We then re-sampled some high-population areas with smaller size increments in order to get a better sample, in some cases sampling in increments of 1 byte. In total, we collected 844,140 files from 304,361 GitHub repositories. Our sample represents 94.09 % of the estimated population of 897,182 files under 29 KB on GitHub written in Shell language and containing the word "alias". The file contents, together with repository metadata, were stored in an SQLite database. After removing duplicate files based on their SHA-1 hash value, our database contains 372,816 unique files from 205,126 repositories.

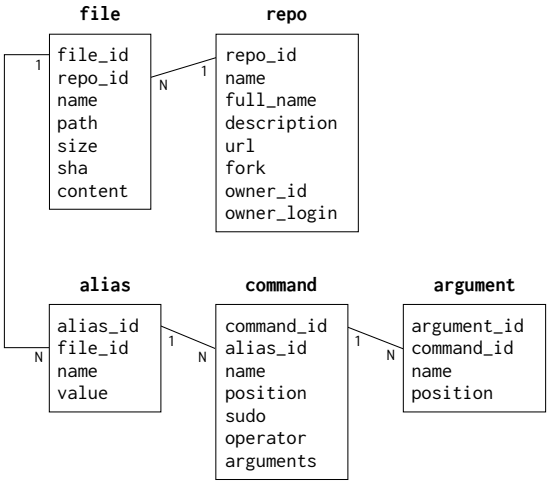


Figure 1: Relational database schema.

3.2 Parsing

After collecting files with potential aliases, we ran a parsing script to find actual alias definitions and decompose them into their constituent parts for analysis. The decomposed aliases are stored in the same SQLite database as the raw file contents to facilitate easy cross-referencing. The database schema is given in Fig. 1.

The parser is a Haskell script that splits each alias definition into alias name and alias value, and tokenizes the value into commands

[4] https://bigquery.cloud.google.com/table/bigquery-public-data:github_repos.files
[5] https://docs.github.com/en/rest/reference/search
[6] GitHub uses the Linguist library to classify code: https://github.com/github/linguist
[7] https://github.com/ipa-lab/github-searcher

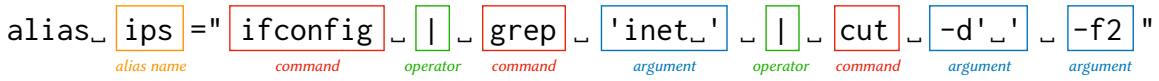


Figure 2: Decomposition of `alias ips="ifconfig | grep 'inet ' | cut -d' ' -f2"`.

Table 1: Distribution of common file names

%	Files	Name Pattern	Aliases	%
14.35	27,870	`*alias*`	612,516	27.79
27.72	53,844	`*bashrc*`	591,396	26.83
22.15	43,011	`*zshrc*`	487,002	22.09
9.42	18,298	`*profile*`	199,009	9.03
1.26	2,455	`git*`	61,248	2.78

and arguments. Commands can be delimited by the shell operators for piping (`|` and `|&`), logical composition (`&&` and `||`), background execution (`&`) and simple chaining (`;`). Arguments are separated by whitespace, but care is taken to handle quoted arguments correctly. For example, `echo "hello world"` is parsed as one command (`echo`) with one argument (`"hello world"`). See Fig. 2 for a more elaborate example.

Beyond quoting, which is defined by the Shell Command Language and thus uniform across all commands, the parser can not make any further considerations as to how arguments are meant to be interpreted. While there are some conventions around command-line argument handling, programs are generally free to do as they wish and there is a wide variety of argument styles in the wild: single-dash short arguments combined with double-dash long-form arguments (e.g., `ls -l -a --color=always`); combined short arguments without a dash (e.g., `tar xvzf archive.tar`); dictionary-style arguments (e.g., `dd if=/dev/zero of=/dev/sda`); subcommands (e.g., `git commit -m "wip"`); and many more. Since the parser can not know the intentions of any command, it simply treats each token as a separate argument. There is one exception: if the command is `sudo`, then its first argument is taken as the real command. For example, `sudo apt-get install` is parsed as the command `apt-get` with argument `install` and the sudo flag set.

After parsing, we ended up with 2,204,199 alias definitions, broken down into 2,534,167 commands and 3,630,423 arguments. Files that did not contain any aliases were removed from the database, as was repository metadata that only referenced files without aliases. 194,218 files from 138,112 repositories, or 52.09 % of the original sample without duplicates, contained aliases.

3.3 Provenance

The majority of aliases in our dataset (85.74 %) originate from common startup scripts, like `.bashrc`, `aliases.zsh` or `.profile` (see Table 1). We found another 2.78 % of aliases originating from scripts related to Git, with file names like `git.plugin.zsh` or `git.bash`. The remaining aliases are more or less evenly distributed among a variety of file names, none of which contributes more than half a percent of aliases, in most cases significantly less. The average number of aliases per file is 11 ± 18, the median is 6.

Table 2: Most common words in repository descriptions

%	Repos	Word in Description	Aliases	%
21.91	30,259	`my`	582,448	26.42
17.00	23,483	`dotfiles`	466,006	21.14
12.75	17,612	`files`	316,963	14.38
6.85	9,466	`configuration`	175,333	7.95
5.33	7,364	`config`	131,430	5.96
4.54	6,269	`personal`	113,885	5.17
4.13	5,707	`linux`	101,747	4.62
3.17	4,385	`bash`	94,739	4.30
3.88	5,353	`scripts`	91,021	4.13
2.06	2,840	`zsh`	74,034	3.36

Table 2 shows the most commonly occurring words in repository descriptions on GitHub (excluding stop words), together with the amount of aliases found in repositories whose descriptions contain at least one of these words. Counting them all together, repositories mentioning any of the words listed in Table 2, in either their description or their repository name, make up 74.48 % of the repositories in our dataset, contributing 82.3 % of all aliases. It is notable that more than half of the repositories in our dataset (51.08 %) have a name that includes the string `dot`, as in `dotfiles`, `dot-files`, `dots`, `mydotfiles`, and so on. Looking at these names and descriptions, we can see a clear bias towards personal configurations and settings management. On average, each repository contributes 16 ± 28 aliases, the median is 8.

3.4 Reproducibility

To enable reproducibility and follow-up studies, we have made all data and our entire tool-chain publicly available. Our dataset (1.45 GB of parsed alias definitions, plus 4.3 GB unparsed file contents and metadata) is available on Zenodo.[8] The parsing script and the executable Jupyter notebooks, containing all SQL queries and additional Python code used during our analysis, are available on GitHub.[9]

4 ANALYSIS

Table 3 shows the most common alias names, commands, and arguments appearing in alias definitions. The most common alias name we found is `ls`, appearing a total number of 83,782 times, which is 3.8 % of all alias definitions. Note that this is `ls` as an *alias name*, a redefinition of the `ls` *command*, which appears 260,156 times (10.27 %). This is a bit less often than `git`, the most common command, which appears in 327,786 aliases (12.93 %). The most common argument, across all commands, is `--color=auto`, appearing 153,931 times (4.24 %)

[8] https://doi.org/10.5281/zenodo.4007049
[9] https://github.com/ipa-lab/shell-alias-analysis

Table 3: Top alias names, commands and arguments.

Name	#	%
`ls`	83,782	3.80
`ll`	62,465	2.83
`grep`	44,479	2.02
`la`	43,760	1.99
`l`	39,539	1.79

Command	#	%
`git`	327,786	12.93
`ls`	260,156	10.27
`cd`	166,632	6.58
`grep`	89,598	3.54
`vim`	46,545	1.84

Argument	#	%
`--color=auto`	153,931	4.24
`-i`	70,640	1.95
`-a`	42,910	1.18
`-l`	39,519	1.09
`-v`	35,295	0.97

Table 4: Top two commands with top arguments and aliases.

	%	Arguments	Aliases (%)
`git`	5.85	`status`	`gs` (54.27), `gst` (19.19)
	3.48		`g` (75.71), `gti` (5.74)
	3.20	`checkout`	`gco` (50.52), `gc` (13.87), `gch` (7.56)
	3.18	`push`	`gp` (46.73), `gps` (9.23), `push` (7.56)
	3.16	`diff`	`gd` (79.89)
	2.86	`pull`	`gpl` (18.30), `gl` (16.59), `gp` (15.07)
	2.78	`branch`	`gb` (73.54), `gbr` (6.57)
	2.71	`add`	`ga` (80.96)
	2.00	`commit`	`gc` (63.16), `gci` (5.33)
	1.96	`commit -m`	`gcm` (31.29), `gc` (25.18), `gm` (7.97)
`ls`	14.45	`--color=auto`	`ls` (99.04)
	8.63	`-A`	`la` (97.61)
	7.80	`-CF`	`l` (98.75)
	6.78	`-alF`	`ll` (97.49)
	5.46	`-l`	`ll` (78.83), `l` (7.91)
	3.75		`l` (27.90), `sl` (21.45)
	2.88	`-G`	`ls` (96.47)
	2.74	`-la`	`ll` (38.42), `la` (26.87), `lla` (12.63)
	2.67	`-a`	`la` (76.94)
	1.92	`-al`	`ll` (49.69), `la` (12.23), `l` (8.49)

Looking at each part of an alias definition in isolation can only get us so far, as arguments only gain meaning in conjunction with commands and alias names can be identical between users, referring to the same command/argument combination, or indeed can overlap, meaning the same alias name is used differently by different users. Table 4 gives a more informative view for the top two commands, `git` and `ls`, showing us the top arguments given with each and the most common alias names by which the command/argument combinations are referred to. Here we can already identify some of the typical alias use cases. Looking at `ls`, we find that aliases are used to redefine the command with a default argument (`ls` → `ls --color=auto`); to shorten a common invocation (`ll` → `ls -alF`); and to correct a spelling mistake (`sl` → `ls`). We also notice that in the case of `git`, most aliases are used for shortening `git` subcommand invocations (e.g. `gd` → `git diff`).

4.1 Inductive Coding

To capture the range of patterns and use cases for which aliases are defined, we analyzed the dataset using inductive coding, a classic technique for qualitative data analysis [13, 47, 52]. Inductive coding is used when conducting exploratory research without prior expectations on themes in the data. The individual data points—in our case, alias definitions—are labelled with descriptive tags which try to capture the essence of the datum for later purposes of categorization. It is an iterative process between theoretical sampling and comparing data within emerging themes, continuing in cycles until no new themes emerge.

Since manually coding the entire dataset is infeasible, we developed our themes by coding a representative sample. For this sample, we gathered the top three most common aliases for the top ten most common arguments for the top 50 commands (cf. Table 4), resulting in 1,381 alias definitions, directly covering 28.77 % percent of the dataset. Additionally, we drew a random sample of 200 alias definitions from the long tail of unique aliases. These are aliases that each occur only once in the entire dataset, making up 27.53 % of all aliases. The commands that occur in this long tail are distributed in roughly the same manner as the commands in the whole dataset, the top commands being `cd`, `git`, `ssh`, `ls`, and `vim`. Unique aliases often contain user-specific file system paths (e.g. `gitbash` → `source /Users/j/mybin/gitsh`), happen to have a unique combination of arguments (e.g. `ls` → `ls -GphF`) or are otherwise highly particular (e.g. `h23` → `history -23000`).

In total, we looked at 1,581 aliases during the coding process. In order to reason about the intent of any particular alias, we had to take the semantics of each command into account, consulting their `man` pages and other forms of documentation.[10] To increase the trustworthiness of our codes, coding was performed independently in parallel by the two authors. After a first iteration, we compared our labels, consolidating different naming conventions. In consecutive iterations, we identified ways of formalizing the emerged categories, i.e. constructing automated mechanisms for classifying alias definitions as belonging to certain categories. The suitability for mechanical classification was an important factor for the viability of any emerging themes. The discussion of these formalizations additionally served to establish a better shared understanding. Ultimately, we reached a saturation point at which further coding and analysis did not lead to further insights.

5 CUSTOMIZATION PRACTICES

We identified nine customization practices among three types of aliases: Shortcuts introduce new names and are often used for *nicknaming commands*, *abbreviating subcommands*, and *bookmarking locations*; Modifications change the semantics of commands by *substituting commands*, *overriding defaults*, *colorizing output*, and *elevating privilege*; and Scripts combine multiple commands,

[10] The website https://explainshell.com has been an indispensable resource.

Table 5: Alias types and customization practices

	#	%
Shortcuts		
Nicknaming Commands	244,872	11.11
Abbreviating Subcommands	194,850	8.84
Bookmarking Locations	321,546	14.59
Modifications		
Substituting Commands	100,564	4.56
Overriding Defaults	319,239	14.48
Colorizing Output	182,623	8.29
Elevating Privilege	93,683	4.25
Scripts		
Transforming Data	74,719	3.39
Chaining Subcommands	22,062	1.00

often for the purposes of *transforming data* or *chaining subcommands*. We developed automated classification methods for each practice, which can be found in our replication package. Table 5 gives a quantitative overview of the prevalence of each of these practices in the dataset. Any alias can be an expression of multiple customization practices at once, and some practices only occur with certain commands. Table 6 breaks down the customization practices by command, counting the number of aliases that a command is involved in (including aliases that redefine the command).

We will now discuss the alias types and customization practices in more detail.

5.1 Shortcuts

The most obvious use of an alias is to give a complex expression a short and/or memorable name. The average length of an alias name is 4.3 characters, whereas the average length of an alias value is 23.7 characters. If we divide the length of an alias value by the length of the alias name, we get the *compression ratio* of the alias. For example, the alias `gs` → `git status` has a compression ratio of 5. Fig. 3 shows the distribution of compression ratios over all aliases in the dataset. The median compression ratio is 4.25, meaning half of all alias values are at least four times as long as their alias names. A compression ratio less than 1 indicates a name that is longer than the value it aliases.

There are 26,055 aliases (1.18 %) with names longer than their values. The two longest alias names we found are from joke definitions. The first is 1,772 characters long and is comprised of the letter 'f' repeated 1,053 times, followed by the letter 'u' repeated 719 times. It is an alias for the `cat` command with a similarly named file as an argument. The second longest alias name is a Swedish compound word of 131 characters,[11] aliasing the `ls` command.

On the other end of the spectrum, an alias named `line` echoes 23,635 dashes, achieving a compression ratio of 5,911, the highest among all aliases. The second highest comes from an alias named `BEEP`, which invokes the Linux `beep` utility 9 times in succession,

[11] Translating, roughly, to northwestern-glacier-artillery-flight-thrust-simulator-plant-equipment-maintenance-follow-up-systems-discussion-posts-preparation-works.

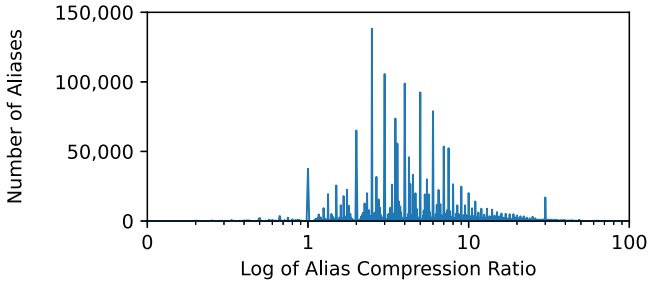


Figure 3: Distribution of alias compression ratios

with a combined 4,471 arguments. When executed, it appears to play Daft Punk's 2001 instrumental single *Aerodynamic*.

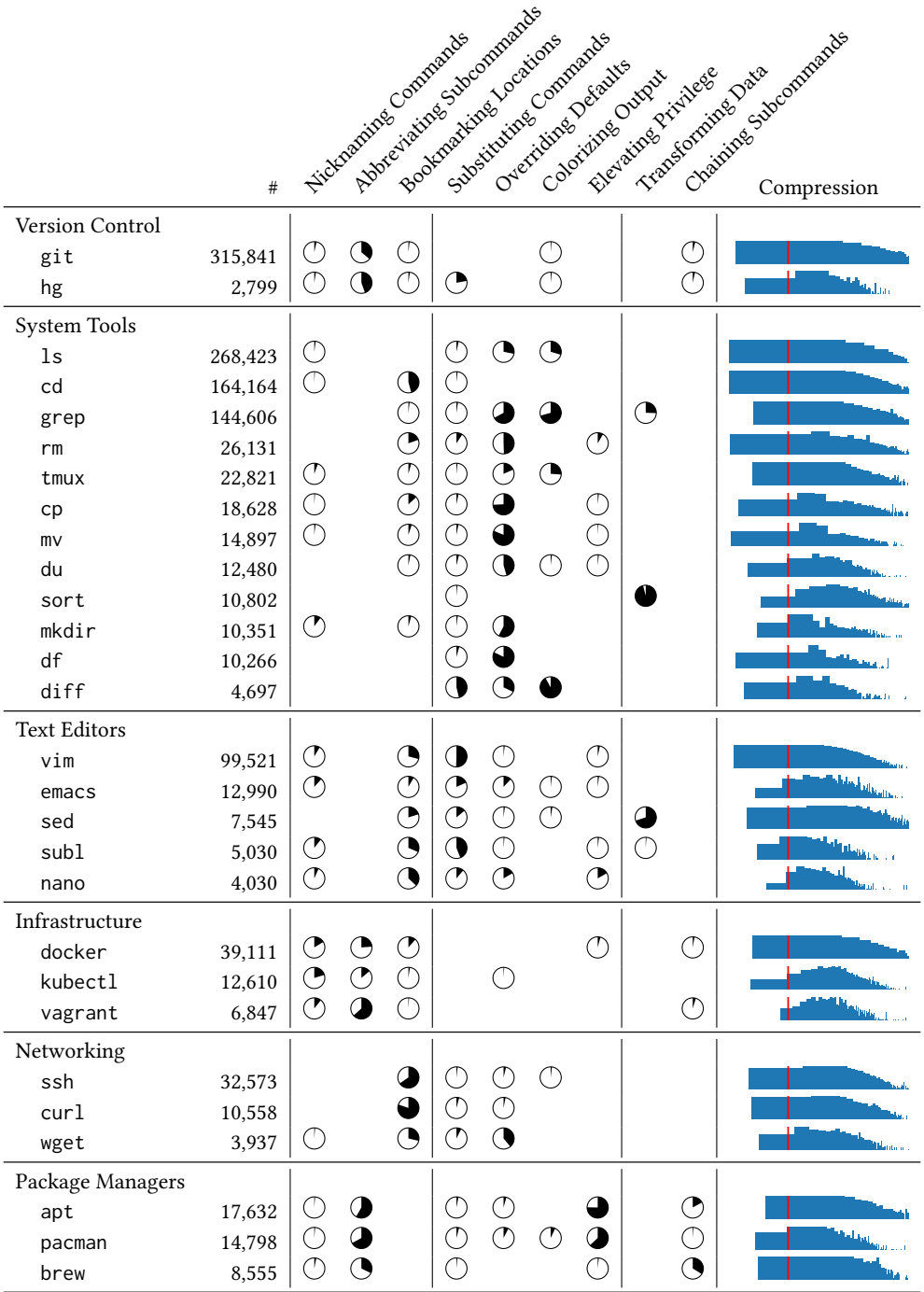
Beyond just compression and expansion of strings, we can see a few distinct customization practices related to naming.

Nicknaming Commands. There are 244,872 aliases in our dataset (11.11 %) that merely give a new name to a command, without adding any arguments, and without the name belonging to a different command (that would be a substitution, see below). The most often occurring nicknames are `g` → `git`, `c` → `clear`, `h` → `history`, and `v` → `vim`. Almost all (93.03 %) of these kinds of aliases introduce a nickname that is shorter than the command they are referring to, and about half (50.58 %) introduce a name that is only one or two characters long.

A special case of nicknaming occurs when the new name is a common misspelling of the command. In this case, the alias acts like an autocorrect mechanism, as in `got` → `git`. To determine instances of these typographical errors, we surveyed and experimented with different string distance measures [35] and decided on using the Damerau-Levenshtein algorithm [9]. We determined empirically that a distance measure of 2 seems like a good threshold to decide whether or not an alias corrects a misspelling. We found 9,195 aliases (0.42 %) that serve as autocorrect rules, most commonly involving transposition (`grpe` → `grep`), case-sensitivity (`Jupyter` → `jupyter`), localization (`pluralise` → `pluralize`), and punctuation (`docker-build` → `docker_build`).

Abbreviating Subcommands. Many commands can operate in different modes, or act as interfaces to a variety of different *subcommands*. The subcommand is commonly specified as the first argument to the command, and takes its own set of arguments and flags. For example, `git push --tags` executes the `push` subcommand of `git` with the `--tags` flag enabled. We identified 67 commands in our dataset that take subcommands, such as `git`, `docker`, or `systemctl`. Noticeably, we found 194,850 aliases (8.84 %) that are purely abbreviations of subcommands, without adding any additional arguments beyond the subcommand. For example, `gs` → `git status` or `gd` → `git diff`. The majority of such subcommand abbreviations (58.5 %) are for `git`, with 113,980 aliases defined purely for abbreviating `git` subcommands, accounting for 36.77 % of all aliases involving `git`. The command with the second-most subcommand abbreviations is the package manager `pacman`, with only 9,918 instances (5.09 % of subcommand abbreviations, but 68.67 % of all aliases involving `pacman`).

Table 6: Customization practices broken down by command. We present a selection of common commands and for each of the nine customization practices show the percentage of occurrences of the command that happen as part of that customization practice, if it is more than 1% of all occurrences of the command. Note that a single command occurrence can be part of multiple customization practices at once. The compression ratio plots are log-log histograms, the red line marks a ratio of 1.



Bookmarking Locations. When an aliased command is called with an argument that references some specific local or remote location, like a file path or domain, the alias acts as a bookmark to that location. For instance, `dl` → `cd ~/Downloads` and `starwars` → `telnet towel.blinkenlights.nl` are both bookmark aliases. To find such bookmarking uses in our dataset, we searched for arguments that are locations, which we take to be any of the following:

- A string containing a forward slash (`/`), indicating a path.
- An IPv4 address, matched by the liberal regular expression `[0-9]+\.[0-9]+\.[0-9]+\.[0-9]+`
- A string containing one of the known top-level domains[12] preceded by a dot (`.`) and followed by a slash (`/`), colon (`:`) or the end of the string.

To avoid false positives, we sampled the top 300 search results according to the above criteria and determined some exclusion patterns. For instance, `/dev/null` is not a location for our purposes. Neither is `origin/master`, and thus an alias like `gm` → `git merge origin/master` does not count as a bookmark. We also exclude aliases that are merely referencing unnamed relative directories (e.g., `../..`).

By our definition, 321,546 aliases (14.59 %) are bookmarks. Of these, 59,931 are remote bookmarks containing URLs or IP addresses (15.92 % of all bookmarks). Bookmarks are used predominantly for file system navigation, and the `cd` command is featured heavily. Most other uses seem to be development related, like starting services such as web servers or databases with pre-defined locations, opening frequently edited files, or outputting logs, as in `onoz` → `cat /var/log/errors.log`

5.2 Modifications

Aliases are not only used syntactically, for naming purposes, but also in ways that change the semantics of certain commands. We found four customization practices related to command modification.

Substituting Commands. When an alias name is identical to the name of a pre-existing command, the alias defines a substitution for that command. A common example is `more` → `less`, replacing a standard Unix utility (`more`) with a more capable but similar command (`less`). This can also be used for subterfuge, as in `emacs` → `vim` (appearing 132 times in our dataset) or indeed `vim` → `emacs` (86 times, alas).

To determine which alias names are also actual command names, we compared them to known Unix commands[13] and a curated sample of commands from our dataset (taking care to not include names that appear in a command position but are actually just other aliases). To determine proper substitutions, we only count aliases whose value does not also include the name of the command (which would point to an overriding alias, see below). We find that 100,564 aliases (4.56 %) are used to substitute one command for another. The top three substitutions are `vi` → `vim`, `vim` → `nvim`, and `vi` → `nvim`.

[12] http://data.iana.org/TLD/tlds-alpha-by-domain.txt
[13] https://en.wikipedia.org/wiki/List_of_Unix_commands
and https://en.wikipedia.org/wiki/List_of_GNU_Core_Utilities_commands

Overriding Defaults. When an alias has the same name as the command it aliases, as in `ls` → `ls -G`, then the alias re-defines the command and effectively overrides its default settings. Any time the command is now executed, it will be with the arguments specified in the alias. There are 319,239 aliases in our dataset (14.48 %) that are used to override defaults in this way. Aliases to override the defaults of the `grep` family of commands (`grep`, `egrep`, `fgrep`) occur 96,970 times, accounting for 4.4 % of all alias definitions (and 68.27 % of all `grep` appearances). The `ls` command is redefined with new defaults 75,374 times, accounting for 3.42 % of all aliases (28.99 % of `ls` appearances).

Looking at the new defaults of these redefined commands, they reveal a variety of user preferences, especially in the diverse long tail, where we find a lot of unique alias definitions and argument combinations. Two areas of customization stand out, however: formatting output and adding safety. The majority of overrides for file system commands (`mv`, `cp`, and `rm`, but also `ln`, for creating symbolic links) enable interactive mode (`-i` and variations), which prompts the user before performing potentially destructive actions. Verbose output (`-v`) also plays a role here, describing exactly what kind of effects a command execution had or will have. Enabling verbosity can also be seen as a kind of output formatting, although much more common is the wish for human-readable output. For example, the alias `df` → `df -h` ensures that the available disk space is displayed in common size units, as opposed to just the raw number of bytes. But by far the most common reason for overriding defaults is to enable colorized output. This behavior is so prevalent that we count it as a customization practice in its own right.

Colorizing Output. Enabling colored output can be done in many different ways: adding an argument (like `less -R` or `grep --color=always`), setting an environment variable (as in `ssh` → `TERM=xterm256color ssh`), running the command through a tool that colorizes its output (like `grcat` or `pygmentize`), or even replacing a command outright (`diff` → `colordiff`). Taking all these varieties into account, more than half of all command redefinitions (57.21 %) enable colored output by default. This amounts to a surprising 182,623 aliases, or 8.29 % percent of all aliases in the dataset. If we extend this count to also include aliases that introduce new names (like `ll` → `ls -l --color=auto`), then more than 10 % of aliases colorize a command's output.

Elevating Privilege. The `sudo` command allows the user to execute another command with superuser privileges. Combining a command with `sudo` is often necessary if the other command needs to modify critical parts of the system. In our dataset, we found 93,683 aliases (4.25 %) in which a command is prefixed with `sudo`. The top `sudo`-prefixed command is the package manager `apt-get`, appearing 10,467 times with `sudo`. Remarkably, these are 89.35 % of all occurrences of `apt-get`. In fact, 72.45 % of all occurrences of the package managers `apt*` (Debian and derivatives; including `apt`, `apt-get`, `apt-cache`, `aptitude`, and `$apt_pref`), `pacman`, `abs` and `aur` (Arch Linux), `yum` (RPM), `dnf` (Fedora), `zypper` (openSUSE), `port` (macOS), and `gem` (Ruby) are together with `sudo`, and these package managers account for 29.1 % of all `sudo` occurrences. Interestingly, the macOS package manager `brew` rarely appears with `sudo` (only 1.07 %), even though it is the third most occurring package manager overall, behind `apt*` and `pacman`.

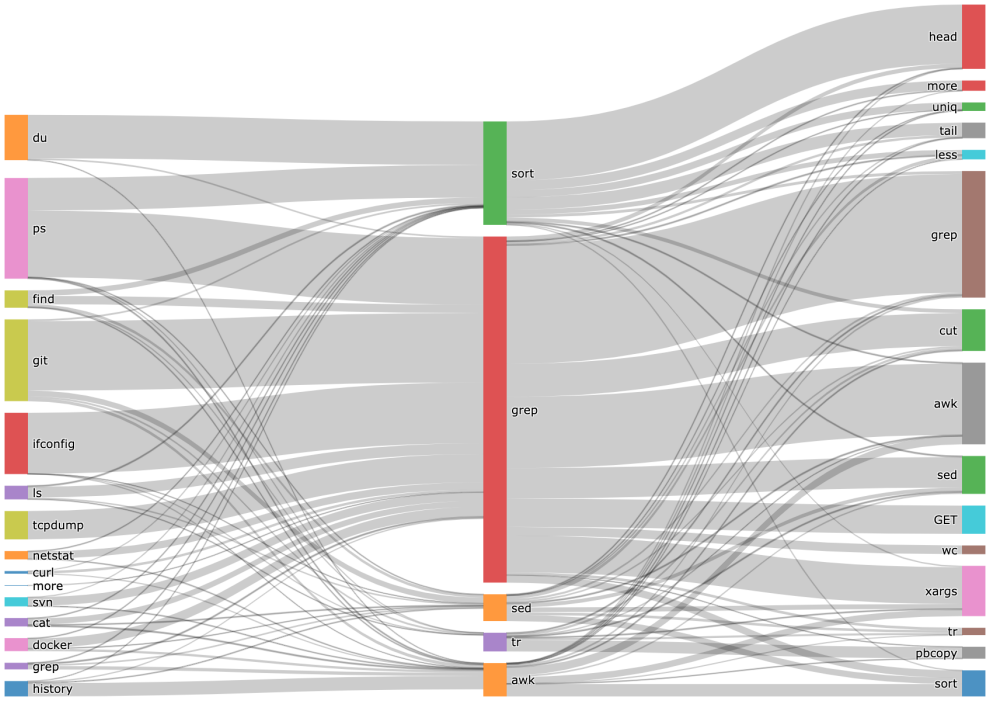


Figure 4: Flow diagram of the top 250 pipelines with three commands that make up at least 10 % of one command's usage

Other commands that more often than not demand elevated privileges are system utilities like `systemctl`, `shutdown`, `lsof` or `mount`.

5.3 Scripts

Aliases that combine multiple commands are basically tiny shell scripts. In our dataset, 204,142 aliases (9.26 %) compose multiple commands. The most popular composition operator is the pipe (`|`), used in 39.66 % percent of alias scripts, followed by the operators for simple chaining (`;`), with 29.61 %, and logical conjunction (`&&`), with 26.88 %. Other operators (`||`, `|&`) appear in only 3.85 % of multi-command aliases.

There are two scripting practices that are of particular interest.

Transforming Data. The pipe (`|`) creates an interface between two otherwise separate programs. It embodies the Unix philosophy of small tools doing one thing well, which can then be connected together to accomplish more complex tasks. There are 74,719 aliases (3.39 %) combining two or more commands using only the pipe operator. The most common command occurring after a pipe, by far, is `grep`, which makes an appearance in almost half of all pipelines (46.16 %), more than three times as often as `xargs` and `sort`. The most common data sources are `ps`, `git`, and `ls`, which are found at the beginning of almost a third (32 %) of all pipelines. Fig. 4 shows a flow diagram of the top pipelines with three commands.

The names of aliases for such pipelines are varied, speaking to the broad range of tasks that can be accomplished by combining various Unix tools. They range from the descriptive, as in `diskspace` → `du -S | sort -n -r | more` or `weather` → `wget -qO - http://wttr.in/ | head -7`, to the very terse, as in `h` → `history | uniq | tail -15` or `lll` → `ls -trlh | less`. Interestingly, aliases with the same name usually describe pipelines with the same general shape (the same commands in the same order), but slightly different argument combinations:

```
lsd → ls -l | grep "^d"
lsd → ls -la | grep ^d
lsd → ls -lGFA --color | grep -i "^d.*/"
lsd → ls -lh | grep --color=never '^d'
```

This highlights the highly personal nature of aliases, each customized for an individual use case.

Chaining Subcommands. An interesting pattern appearing in alias scripts are chains of subcommand invocations. For example, the package manager `brew` has a subcommand `update`, for updating the package database, and a subcommand `upgrade`, for upgrading previously installed packages to the latest available versions. 28.08 % of all aliases involving the `brew` command contain the composition `brew update && brew upgrade` (sometimes with `;` instead of `&&`), with alias names like `update`, `brewup`, `bup`, etc. This pattern of repeated subcommand invocations can be found in 22,062 aliases (1 %), and it is most prevalent among package managers, like `brew`, `apt-get`, `npm` or `gem`, mostly for the same purpose as above.

The command with the highest absolute number of aliases showing this pattern is `git`, however, with 12,063 occurrences (3.89 % of all aliases using `git`). Here, the uses are more varied, e.g., `commit` → `git add . && git commit -m`, or `gitpull` → `git stash && git pull && git stash pop`, or indeed `whoops` → `git reset --hard && git clean -df`.

6 IMPLICATIONS

Through our large-scale analysis of the collective knowledge of shell customization via aliases, we gained insight into practices detailing how users customize their command-line interface. Based on our observations, we outline discussion points that go beyond single customization practices and identify implications that can address shortcomings in command-line usability and tie them to existing user experience research. Further, while our presented findings already give us an understanding of customization practices over many different kinds of commands, we view our collected dataset as a playground for fine-grained discovery that can benefit researchers, tool builders, and command-line users.

6.1 Learning Repair Rules

The complexity of commands and arguments can cause users to introduce errors when working in a command-line interface. Figuring out specifically how to fix these errors is often a convoluted process. A popular open source project that attempts to navigate this issue[14] uses a set of rules to suggest possible error corrections for commands. While these rules are all hard-coded, we envision leveraging the global wisdom of customizations in our large-scale dataset to learn rules that form the basis for different kinds of suggestions. This is in line with visions of integrating collective intelligence in software development [6], in particular work in leveraging emergent behavior from corpora [14] that we can codify based on our customization data. We can also see approaches similar to work on learning code completions from examples [7], with our dataset of alias definitions serving as an oracle for an automatic software repair system [34] in the domain of shell commands. Using our dataset of known-good command invocations, it should be possible to train a statistical language model for command repair, akin to related work in code synthesis [44].

As an example, take the following erroneous invocation:

```
$ apt-get install vim
E: Could not open lock file /var/lib/dpkg/lock - open
↪  (13: Permission denied)
E: Unable to lock the administration directory
↪  (/var/lib/dpkg/), are you root?
```

Without having to consult a hard-coded rule involving knowledge about `apt-get`, or even looking at the specific error that is produced, a command repair system trained on our dataset of alias definitions could easily suggest the correct fix: `sudo apt-get install vim`. It is reasonable to assume that this could be inferred as the correct invocation, because in aliases the command sequence `apt-get install` occurs almost exclusively pre-fixed with `sudo`.

As another example, the following error is caused by the wrong order of arguments to the `systemctl` command:

```
$ systemctl docker status
Unknown command verb docker.
```

The correct invocation is `systemctl status docker`. It is again very plausible that a repair rule for this type of error could be learned from our dataset, based on the prevalence of aliases containing the command `systemctl` together with an argument `status` that occurs in first position, indicating the latent knowledge that `status` is in fact a subcommand of `systemctl`.

6.2 Discovering Workflows

Following a different thread of leveraging emergent practices, we can also see how our dataset would enable a world beyond only trying to fix immediate errors, by providing usage hints that could introduce users to common parameters and workflows. For example, as soon as a user tries to `sort` the output of the `ps` command, the alias `mem10` → `ps auxf | sort -nr -k 4 | head -10` can serve as a suggestion for the complex but common data transformation that results in showing the ten most memory-intensive processes. Similarly, in the practice of *chaining subcommands* we can clearly see the prevalence of object protocols [5], which are implicit rules determining the order in which commands have to be executed. We can improve usability by enabling the discovery of these implicit rules and by exposing the dependency structure based on our customization data. For instance, if executing `brew upgrade` results in a failure, we can suggest using `brew update && brew upgrade` instead, based on the patterns in our dataset (cf. Section 6.1).

Our findings can also contribute to recent work on the parallelization and distribution of shell scripts. Systems like PaSh [54] and POSH [43] rely on manual annotation of commands and their arguments to effectively parallelize shell scripts. Our data can help focus these annotation efforts by informing the developers of these systems about which groups of commands and arguments are most frequently used together. The KumQuat system [55] leverages program synthesis techniques to search a large space of candidate solution to synthesize parallel shell scripts. The collective knowledge present in alias definitions can guide this search and justify certain intuitions about the latent data parallelism in Unix pipelines [23]. For example, while a parallel version of the `comm` command for comparing sorted files line-by-line is not synthesizable in general, it becomes trivially parallelizable if each of its input lines is known to be unique. Evidence that this indeed the common case can be found in our dataset, where 41.29 % of all occurrences of `comm` follow `sort | uniq` or `sort -u`, and the remainder mostly have unique data sources as input, like `pacman -Qeq`.

6.3 Uncovering Conceptual Design Flaws

Customization can also be an indicator for problems in the underlying conceptual design, manifesting as usability frustrations that require adaptation by the user. In their analysis of Git, Perez De Rosso and Jackson [38, 39] describe a number of flaws and operational misfits arising from the conceptual design of the software. The frustrations experienced by users because of these design flaws are evident based on the alias definitions in our dataset.

For example, the difficulties some Git users have with the concept of *staging* can be seen in aliases that ensure untracked files are included in a commit by explicitly adding them beforehand, like `commit` → `git add . && git commit -m` or `gac` → `git add --all && git commit`.[15] Another frustration is having to use `git stash` to temporarily save uncommitted changes and clean the working directory in order to avoid conflicts when using other

[14] https://github.com/nvbn/thefuck

[15] See "Just Let Me Commit!" in [38] and "Incoherent Commit" in [39]. Confusingly, `git commit -a`, while performing an implicit add, does not include untracked files.

Git commands. Stashing in itself has no higher purpose in version control, it merely exists as a concept to work around limitations in Git.[16] This can be seen in aliases like `gspull` → `git stash && git pull && git stash pop`, which defines a new type of pull command that stashes away ongoing work before pulling in remote changes and finally re-applying the stashed work. The same problem happens when switching branches, hence aliases like `gsc` → `git stash && git checkout $1 && git stash pop`.

Church et al. [8] found that version control systems are generally perceived as being risky to use, and sought explanations for this impression via an analysis of Git using a framework of cognitive dimensions [17]. One of the dimensions that dominate the command-line interface of Git is *Hidden Dependencies*. The are many hidden dependencies in Git, a prominent one being the dependency between the local branch and the remote repository. This is revealed by alias definitions like `gitstatus` → `git remote update && git status`. Unless one first manually updates Git's local information about remote branches, the command `git status` will happily report that the local branch is up-to-date with respect to its remote origin, even if the remote repository is in fact many commits ahead.

We want to emphasize that we are not suggesting that large-scale quantitative data of customization practices can replace qualitative analysis, but rather that the corpus we provide, together with our findings, can support exploration and provide new insights for usability research. Alias definitions can provide evidence for analytic theories based on cognitive or conceptual models of software use, because they codify workarounds for common annoyances and other customizations based in every-day use. According to a recent need-finding study by Zhang et al. [58], API designers have a strong desire to know more about users' mental models, and wish to validate design hypotheses with examples of real-world API usage. Existing techniques for mining API usage fall short in this respect, and the study highlights the importance of, among other things, looking at how users deal with unanticipated corner cases and how they apply workarounds. We suspect makers of command-line software are in a similar situation as API designers and could similarly benefit from community usage data that highlights gaps between interface design and users' expectations.

6.4 Contextual Defaults

Choosing proper defaults in user interfaces is a pillar of user experience design [36]. The fact that 14.48 % of the customizations in our dataset are for *overriding defaults* suggests that, at least for some groups of users, the default settings of their tools could be improved. We see *overriding defaults* not necessarily as an indictment of the involved commands, but rather as an indication that the assumed user context does not in all cases match the actual usage profile. This can be the case if the tool assumes a different execution environment than the one it is ultimately used in, e.g. personal notebook vs cloud deployment (where an alias like `java` → `java -ea -server` ensures that Java programs are always run on a server-optimized virtual machine) or interactive terminal vs shell script use (cf. Section 6.5), or if the tool assumes a certain type of user with different needs than the actual user.

Indeed, the variety of different defaults in the data indicate what we call *contextual* defaults, where context could be a reflection of the level of expertise of a command-line user, or a certain persona (e.g., system administrator, data scientist, or software engineer). For example, the top default alias for the `ffmpeg` command is `ffmpeg` → `ffmpeg -hide_banner`, suppressing verbose default output that can be confusing for newcomers but is helpful for the tool developers when providing support and locating errors.[17] We could imagine providing different sets of defaults to different users, effectively alias starter packs, generated from our data. We see parallels to work that investigates contextual preferences and personalization in information systems [12, 51] and privacy research [2, 56].

6.5 Interactivity vs Scripting

The first "modern" command line, the Bourne shell from 1977, had two primary goals: to provide an interactive command interpreter, and at the same time serve as a scripting system [27]. There is a natural tension between these two goals, which becomes evident when users are *overriding defaults* with aliases like `mv` → `mv -i`. Here, the `mv` command is redefined to always run interactively, prompting the user at critical points, i.e. before overwriting existing files. The default operating mode of `mv`, and most other commands, is to assume that the user is aware of and okay with the possible consequences of running it—and that they have not made any mistakes in its invocation. This is of course a much more useful assumption in a scripting context.

The bias of most command-line tools towards scripting is also evident in their output, which is usually minimal and not tailored for human ease-of-use. We can see this in aliases like `mount` → `mount | column -t`, which aligns the output of the `mount` command for easier reading, or `df` → `df -h` or `ll` → `ls -lh`, which change the default output of these commands so that file sizes are not shown simply in bytes but rather in much more practical common units like megabytes. The high prevalence of aliases for *colorizing output* (e.g. `grep` → `grep --color=auto`) is also notable, as color only makes sense in an interactive context. In terminals, colorful text is achieved by inserting ANSI escape codes into the text stream. This is a hindrance for scripts, but tools could easily detect whether they are run in an interactive terminal or as part of a script and adjust their output accordingly.

Note that the tension between interactivity and scripting is not the same as the divide between "casual" and "power" users. Experts are experiencing the same frustrations as amateurs when using the shell interactively. Recently, there has been a growing movement that sees today's command line as a *human-first* text-based UI, rather than a *machine-first* scripting platform [42]. This new generation of command-line users and tool authors embrace the Unix philosophy with its core tenet of simple tools that can be composed well together [45], but they want to modernize those tools to fit current environments, with a more humanistic approach to

[16] See "I Just Want To Switch Branches!" in [38], and especially "Unmotivated Stash" and "Branch Coupling" in [39].

[17] Coincidentally, we also found a ticket in the project's issue tracker requesting this top most default argument from our dataset to become the default option for the command: https://trac.ffmpeg.org/ticket/7211

their interaction design.[18] Emphasizing the conversational nature of the command line, they highlight the need for features such as error correction (cf. Section 6.1) or command suggestions (cf. Section 6.2), and confirming potentially destructive actions before they are executed. They see human-readable output as paramount and suggest tools should be more aware of their environment (cf. Section 6.4).

7 THREATS TO VALIDITY

We review potential limitations of our study as threats to validity. First, our sample might not be representative. Our dataset only includes aliases by people who publicly shared their dotfiles, we only collected from GitHub, and our sample does not include forks. Nevertheless, our dataset is very exhaustive, as we were able to sample 94.09 % of the estimated population of Shell files containing aliases on GitHub. And while mining GitHub can be fraught with perils [28], we specifically sought out personal repositories, side-stepping many of the typical issues with mining GitHub for software projects.

Second, our parser might not be sophisticated enough to recognize complex real-world aliases or cope with minute platform differences. To mitigate this threat, we ran multiple sanity checks and tested the parser on some hairy examples from the dataset. We did not detect any significant mis-parses and think that we have covered the majority of relevant cases. The raw unparsed database is available in our replication package.

Third, aliases might not reflect intent as much as we assume. En-masse copy-pasting of aliases by users, without them knowing exactly what they are copying, is certainly a realistic scenario. System distributions and configuration frameworks like *ohmyzsh* ship with numerous aliases by default or as part of easily enabled plugins. Users might not even be aware of the aliases they have on their system. We mitigate this concern by removing all duplicate files from our dataset that would indicate sheer copy/pasting.

Fourth, we might not actually be able to see the true user intent, if it exists, as quantitative measures might hide a long tail of minor variations and individual user preference. Conclusions about common aliases or selected subsets might not be generalizable. To mitigate these summarizing effects, we established customization practices as a vehicle to take a deeper dive into the details of certain alias usage. Since we sampled almost the whole available population, we are confident in the strength of our data and the conclusions we can draw from particular instances. Our replication package includes our whole toolchain and all alias data in a relational format ready for further analysis.

8 RELATED WORK

Related research in the broader context of our work has been conducted on understanding common practices in the software engineering community based on public online data, on software configuration in general, on the use of command-line interfaces and how to improve them, and on the shell as a programming language for both scripting and interactive use.

Empirical studies similar to ours, looking at community knowledge in software engineering to understand practices and distill insights, have been conducted in related domains: Zhong and Su [60] study real-world bug fixes in Java projects to help guide automatic program repair; Yang et al. [57] mine Stack Overflow posts and GitHub repositories to find out how programmers use and adapt copy-pasted code snippets in open-source projects, while Baltes and Diehl [3] investigate to what extent such snippets are copied without proper attribution; Prana et al. [41] conduct a qualitative study to categorize the content of GitHub README files and build an automated classifier to label README sections, easing information discovery; Barnaby et al. [4] present a tool that mines code bases for idiomatic usage examples of API methods.

In the context of software configuration, Sayagh et al. [49] surveyed experts and the literature to identify a number of challenges and recommendations related to configuration practices. Our work reflects some of their findings, insofar as shell aliases are a form of personal configuration that can interact with—and counteract—other system configurations. For example, selecting good out-of-the-box default values is seen as an important issue by experts, and aliases are indeed often used to *override defaults*. Related to our implications on contextual defaults (Section 6.4), Zheng et al. [59] present MassConf, a system that proposes optimal software configurations based on a user's environment and existing configurations. Adjacent work in configuration mining includes the ConfigMiner tool by Sayagh and Hassan [48], which identifies appropriate configuration options based on related StackOverflow questions.

The earliest study we found on the use of command-line interfaces was by Greenberg [21], who collected four months of continuous real-life use of the Unix `csh` shell from 168 users. The data was used in a follow up study to analyze the use of interactive systems by examining the frequency of command invocations for different groups of users [22]. In later work, Davison and Hirsh [11] use probabilistic action modeling to predict user action sequences based on the same dataset. Korvemaker and Greiner [30] similarly predict future action sequences in command lines, but condition on actions of the particular user group with the goal of enabling adaptive user interfaces. Other work in the context of adaptive user interfaces by Jacobs and Blockeel [26] uses association rule learning on the shell logs to produce scripts to automate common task sequences. Khosmood et al. [29] use the same corpus and two additional, more recent, corpora to learn a model that can identify user profiles based on their command-line behavior. Bespoke [53] is a system that synthesizes specialized graphical user interfaces (GUIs) based on command usage. Our work can be viewed as an input to this system that passes common shell workflows in aliases to be generated as GUIs.

There has been other work on enhancing user experience in command-line interfaces. NoFAQ [10] provides repair suggestions for failed shell invocations based on a model learned from a curated set of fix patterns. NL2Bash [31] implements a system that translates natural language phrases in English to shell commands. Recent work by Greenberg [18] has been looking into understanding the

[18] It should be noted that Ritchie and Thompson [46], in a paper that pre-dates the invention of the Bourne shell, explicitly highlight the interactivity of the Unix system over the batch-processing nature of its predecessors. Today's notion of interactivity is of course more advanced, and it is now the classic Unix systems with their shell scripts that evoke an atmosphere of "batch processing."

POSIX shell as a programming language. More specifically, understanding word expansion in the shell to support interactivity [20] and concurrency [19].

9 CONCLUSION

We report on a large-scale exploratory study on how command-line users customize user experience by defining shell aliases. Through inductive coding, nine customization practices emerged from our dataset of collective customization knowledge mined from GitHub, providing insight on the characteristics of command-line use. Based on our results, we discuss and formulate a set of implications for command-line tool developers, researchers, and the shell as an interactive environment for experts. We enable further analysis and a basis for learning applications based on our extensive curated dataset.

Aliases often redefine commands with other default arguments, which is a potential indicator for usability problems in these tools. However, we have to also be aware that defaults can be highly contextual depending on user profiles (e.g., expertise level) and environment (e.g., scripting vs. interactive use). We also see our dataset and results as a rich source for learning norms with respect to repair rules, data flows, and descriptive names for complex command structures. We provide a comprehensive replication package and see potential for future work based on our dataset and analyses.

REFERENCES

[1] Mayank Agarwal, Jorge J. Barroso, Tathagata Chakraborti, Eli M. Dow, Kshitij P. Fadnis, Borja Godoy, and Kartik Talamadupula. 2020. CLAI: A Platform for AI Skills on the Command Line. arXiv:2002.00762 [cs.HC].

[2] Md Zulfikar Alom, Barbara Carminati, and Elena Ferrari. 2019. Helping Users Managing Context-Based Privacy Preferences. In *2019 IEEE International Conference on Services Computing (SCC)* (Milan, Italy). IEEE, 100–107.

[3] Sebastian Baltes and Stephan Diehl. 2019. Usage and Attribution of Stack Overflow Code Snippets in GitHub Projects. *Empirical Softw. Engg.* 24, 3 (June 2019), 1259–1295. https://doi.org/10.1007/s10664-018-9650-5

[4] Celeste Barnaby, Koushik Sen, Tianyi Zhang, Elena Glassman, and Satish Chandra. 2020. Exempla Gratis (E.G.): Code Examples for Free. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (Virtual Event, USA) *(ESEC/FSE 2020)*. Association for Computing Machinery, New York, NY, USA, 1353–1364. https://doi.org/10.1145/3368089.3417052

[5] Nels E. Beckman, Duri Kim, and Jonathan Aldrich. 2011. An Empirical Study of Object Protocols in the Wild. In *Proceedings of the 25th European Conference on Object-Oriented Programming* (Lancaster, UK) *(ECOOP'11)*. Springer-Verlag, Berlin, Heidelberg, 2–26.

[6] Marcel Bruch, Eric Bodden, Martin Monperrus, and Mira Mezini. 2010. IDE 2.0: Collective Intelligence in Software Development. In *Proceedings of the 2010 FSE/SDP Workshop on the Future of Software Engineering Research* (Santa Fe, United States). https://doi.org/10.1145/1882362.1882374

[7] Marcel Bruch, Martin Monperrus, and Mira Mezini. 2009. Learning from Examples to Improve Code Completion Systems. In *Proceedings of the 7th joint meeting of the European Software Engineering Conference and the ACM Symposium on the Foundations of Software Engineering* (Amsterdam, Netherlands). https://doi.org/10.1145/1595696.1595728

[8] Luke Church, Emma Söderberg, and Elayabharath Elango. 2014. A case of computational thinking: The subtle effect of hidden dependencies on the user experience of version control. In *Proceedings of the 25th Annual Workshop of the Psychology of Programming Interest Group* (Brighton, UK). PPIG, 123–128. http://www.sussex.ac.uk/Users/bend/ppig2014/13Church-Soderberg-Elango-PPIG2014.pdf

[9] Fred J Damerau. 1964. A technique for computer detection and correction of spelling errors. *Commun. ACM* 7, 3 (1964), 171–176.

[10] Loris D'Antoni, Rishabh Singh, and Michael Vaughn. 2017. NoFAQ: synthesizing command repairs from examples. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*. ACM, 582–592.

[11] Brian D Davison and Haym Hirsh. 1998. Predicting sequences of user actions. In *Notes of the AAAI/ICML 1998 Workshop on Predicting the Future: AI Approaches to Time-Series Analysis*. 5–12.

[12] Sandra De Amo, Mouhamadou Saliou Diallo, Cheikh Talibouya Diop, Arnaud Giacometti, Dominique Li, and Arnaud Soulet. 2015. Contextual preference mining for user profile construction. *Information Systems* 49 (2015), 182–199.

[13] Ian Dey. 2003. *Qualitative data analysis: A user friendly guide for social scientists*. Routledge.

[14] Ethan Fast, Daniel Steffee, Lucy Wang, Joel R. Brandt, and Michael S. Bernstein. 2014. Emergent, Crowd-Scale Programming Practice in the IDE. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 2491–2500. https://doi.org/10.1145/2556288.2556998

[15] Ishaan Gandhi and Anshula Gandhi. 2020. Lightening the Cognitive Load of Shell Programming.

[16] Georgios Gousios, Bogdan Vasilescu, Alexander Serebrenik, and Andy Zaidman. 2014. Lean GHTorrent: GitHub Data on Demand. In *Proceedings of the 11th Working Conference on Mining Software Repositories* (Hyderabad, India) *(MSR 2014)*. Association for Computing Machinery, New York, NY, USA, 384–387. https://doi.org/10.1145/2597073.2597126

[17] Thomas R. G. Green and Marian Petre. 1996. Usability analysis of visual programming environments: a 'cognitive dimensions' framework. *Journal of Visual Languages & Computing* 7, 2 (1996), 131–174.

[18] Michael Greenberg. 2017. Understanding the POSIX shell as a programming language. *Off the Beaten Track* (2017).

[19] Michael Greenberg. 2018. The POSIX shell is an interactive DSL for concurrency. *DSLDI* (2018).

[20] Michael Greenberg. 2018. Word expansion supports POSIX shell interactivity. In *Conference Companion of the 2nd International Conference on Art, Science, and Engineering of Programming*. ACM, 153–160.

[21] Saul Greenberg. 1988. Using UNIX: Collected traces of 168 users.

[22] Saul Greenberg and Ian H Witten. 1988. Directing the user interface: how people use command-based computer systems. *IFAC Proceedings Volumes* 21, 5 (1988), 349–355.

[23] Shivam Handa, Konstantinos Kallas, Nikos Vasilakis, and Martin Rinard. 2021. An Order-Aware Dataflow Model for Parallel Unix Pipelines. arXiv:2012.15422 [cs.PL].

[24] Pengpeng Hou, Heng Zhang, Yanjun Wu, Jiageng Yu, Yuxia Miao, and Yang Tai. 2021. FindCmd: A personalised command retrieval tool. *IET Software* 15, 2 (2021), 161–173.

[25] IEEE and The Open Group. 2018. The Open Group Base Specifications Issue 7 (IEEE Std 1003.1-2017).

[26] Nico Jacobs and Hendrik Blockeel. 2001. From shell logs to shell scripts. In *International Conference on Inductive Logic Programming*. Springer, 80–90.

[27] M. Tim Jones. 2011. Evolution of shells in Linux. https://developer.ibm.com/tutorials/l-linux-shells

[28] Eirini Kalliamvakou, Georgios Gousios, Kelly Blincoe, Leif Singer, Daniel M. German, and Daniela Damian. 2014. The Promises and Perils of Mining GitHub. In *Proceedings of the 11th Working Conference on Mining Software Repositories* (Hyderabad, India) *(MSR 2014)*. Association for Computing Machinery, New York, NY, USA, 92–101. https://doi.org/10.1145/2597073.2597074

[29] Foaad Khosmood, Phillip L Nico, and Jonathan Woolery. 2014. User identification through command history analysis. In *2014 IEEE Symposium on Computational Intelligence in Cyber Security (CICS)*. IEEE, 1–7.

[30] Benjamin Korvemaker and Russell Greiner. 2000. Predicting UNIX command lines: Adjusting to user patterns. In *AAAI/IAAI*. 230–235.

[31] Xi Victoria Lin, Chenglong Wang, Luke Zettlemoyer, and Michael D. Ernst. 2018. NL2Bash: A Corpus and Semantic Parser for Natural Language Interface to the Linux Operating System. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (Miyazaki, Japan). European Language Resources Association (ELRA), 3107–3118.

[32] Kim Mens and Angela Lozano. 2014. Source code-based recommendation systems. In *Recommendation Systems in Software Engineering*. Springer, Berlin, Heidelberg, 93–130.

[33] Thaís Mombach and Marco Tulio Valente. 2018. GitHub REST API vs GHTorrent vs GitHub Archive: A Comparative Study.

[34] Martin Monperrus. 2018. Automatic Software Repair: A Bibliography. *ACM Comput. Surv.* 51, 1, Article 17 (Jan. 2018), 24 pages. https://doi.org/10.1145/3105906

[35] Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM computing surveys (CSUR)* 33, 1 (2001), 31–88.

[36] Jakob Nielsen. 2005. The power of defaults. https://www.nngroup.com/articles/the-power-of-defaults

[37] Jakob Nielsen. 2005. Ten usability heuristics. http://www.nngroup.com/articles/ten-usability-heuristics

[38] Santiago Perez De Rosso and Daniel Jackson. 2013. What's Wrong with Git? A Conceptual Design Analysis. In *Proceedings of the 2013 ACM International Symposium on New Ideas, New Paradigms, and Reflections on Programming & Software* (Indianapolis, Indiana, USA) *(Onward! 2013)*. Association for Computing Machinery, New York, NY, USA, 37–52. https://doi.org/10.1145/2509578.2509584

[39] Santiago Perez De Rosso and Daniel Jackson. 2016. Purposes, Concepts, Misfits, and a Redesign of Git. In *Proceedings of the 2016 ACM SIGPLAN International Conference on Object-Oriented Programming, Systems, Languages, and Applications* (Amsterdam, Netherlands) *(OOPSLA 2016)*. Association for Computing Machinery, New York, NY, USA, 292–310. https://doi.org/10.1145/2983990.2984018

[40] Louis Pouzin. 1965. The SHELL: A Global Tool for Calling and Chaining Procedures in the System. Multics Design Notebook, Section IV. https://people.csail.mit.edu/saltzer/Multics/Multics-Documents/MDN/MDN-4.pdf

[41] Gede Artha Azriadi Prana, Christoph Treude, Thung Ferdian, Thushari Atapattu, and David Lo. 2019. Categorizing the Content of GitHub README Files. *Empirical Softw. Engg.* 24, 3 (June 2019), 1296–1327. https://doi.org/10.1007/s10664-018-9660-3

[42] Aanand Prasad, Ben Firshman, Carl Tashian, and Eva Parish. 2021. Command Line Interface Guidelines. https://clig.dev

[43] Deepti Raghavan, Sadjad Fouladi, Philip Levis, and Matei Zaharia. 2020. POSH: A Data-Aware Shell. In *2020 USENIX Annual Technical Conference (USENIX ATC 20)*. 617–631.

[44] Veselin Raychev, Martin Vechev, and Eran Yahav. 2014. Code Completion with Statistical Language Models. In *Proceedings of the 35th ACM SIGPLAN Conference on Programming Language Design and Implementation* (Edinburgh, United Kingdom) *(PLDI '14)*. Association for Computing Machinery, New York, NY, USA, 419–428. https://doi.org/10.1145/2594291.2594321

[45] Eric S. Raymond. 2003. *The Art of Unix Programming*. Addison-Wesley Professional.

[46] Dennis Ritchie and Ken Thompson. 1974. The UNIX time-sharing system. *Commun. ACM* 17, 7 (1974), 365–375. https://doi.org/10.1145/361011.361061

[47] Johnny Saldaña. 2016. *The Coding Manual for Qualitative Researchers* (3 ed.). SAGE.

[48] Mohammed Sayagh and Ahmed E. Hassan. 2020. ConfigMiner: Identifying the Appropriate Configuration Options for Config-related User Questions by Mining Online Forums. *IEEE Transactions on Software Engineering* (2020), 1–1. https://doi.org/10.1109/TSE.2020.2973997

[49] Mohammed Sayagh, Noureddine Kerzazi, Bram Adams, and Fabio Petrillo. 2020. Software Configuration Engineering in Practice - Interviews, Survey, and Systematic Literature Review. *IEEE Transactions on Software Engineering* 46, 6 (June 2020), 646–673. https://doi.org/10.1109/tse.2018.2867847

[50] Christian Lee Seibold. 2020. Shell History: Unix. https://paled.handmade.network/blogs/p/7462-shell_history_-_unix

[51] Kostas Stefanidis, Evaggelia Pitoura, and Panos Vassiliadis. 2011. Managing contextual preferences. *Information Systems* 36, 8 (2011), 1158–1180.

[52] David R Thomas. 2006. A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation* 27, 2 (2006), 237–246.

[53] Priyan Vaithilingam and Philip J Guo. 2019. Bespoke: Interactively Synthesizing Custom GUIs from Command-Line Applications By Demonstration. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 563–576.

[54] Nikos Vasilakis, Konstantinos Kallas, Konstantinos Mamouras, Achilles Benetopoulos, and Lazar Cvetković. 2021. PaSh: light-touch data-parallel shell processing. *Proceedings of the Sixteenth European Conference on Computer Systems* (Apr 2021). https://doi.org/10.1145/3447786.3456228

[55] Nikos Vasilakis, Jiasi Shen, and Martin Rinard. 2020. Automatic Synthesis of Parallel and Distributed Unix Commands with KumQuat. arXiv:2012.15443 [cs.PL].

[56] Primal Wijesekera, Joel Reardon, Irwin Reyes, Lynn Tsai, Jung-Wei Chen, Nathan Good, David Wagner, Konstantin Beznosov, and Serge Egelman. 2018. *Contextualizing Privacy Decisions for Better Prediction (and Protection)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173842

[57] Di Yang, Pedro Martins, Vaibhav Saini, and Cristina Lopes. 2017. Stack Overflow in Github: Any Snippets There?. In *Proceedings of the 14th International Conference on Mining Software Repositories* (Buenos Aires, Argentina) *(MSR '17)*. IEEE Press, 280–290. https://doi.org/10.1109/MSR.2017.13

[58] Tianyi Zhang, Björn Hartmann, Miryung Kim, and Elena L. Glassman. 2020. Enabling Data-Driven API Design with Community Usage Data: A Need-Finding Study. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376382

[59] Wei Zheng, Ricardo Bianchini, and Thu D. Nguyen. 2011. MassConf: Automatic Configuration Tuning by Leveraging User Community Information. In *Proceedings of the 2nd ACM/SPEC International Conference on Performance Engineering* (Karlsruhe, Germany) *(ICPE '11)*. Association for Computing Machinery, New York, NY, USA, 283–288. https://doi.org/10.1145/1958746.1958786

[60] Hao Zhong and Zhendong Su. 2015. An Empirical Study on Real Bug Fixes. In *Proceedings of the 37th International Conference on Software Engineering - Volume 1* (Florence, Italy) *(ICSE '15)*. IEEE Press, 913–923.