Apache Software Foundation Incubator Project Sustainability Dataset

Likang Yin
UC Davis
lkyin@ucdavis.edu

Zhiyuan Zhang
UC Davis
zyzh@ucdavis.edu

Qi Xuan
Zhejiang University of Technology
xuanqi@zjut.edu.cn

Vladimir Filkov
UC Davis
vfilkov@ucdavis.edu

***Abstract*—Open Source Software success and sustainability is critically important for the digital infrastructure as OSS is used broadly and yet 83+% of such projects fail. To increase chances of success many projects join established software communities, e.g. the Apache Software Foundation (ASF), with clearly established rules and support. Specifically at ASF, projects that strive to join ASF and are at a nascent development stage are digitally housed in the ASF incubator (ASFI), which provides a mature governance environment and expert help toward long-term sustainability. Projects in ASFI eventually conclude their incubation by graduating, if successful on the path to sustainability. Otherwise, they get retired. In ASF, digital traces of developer activities for projects in ASFI are publicly available, together with monthly project status.**

Here we present a longitudinal dataset of developer coding and communication activities of 269 projects from the Apache Software Foundation Incubator (ASFI). Each project in ASFI is evaluated while in incubation and is eventually "graduated" or "retired", a label indicating the project sustainability promise with respect to their technical development and community diversity. This extrinsically labeled dataset offers heretofore unavailable sustainability data of OSS project development under ASF regulations and governance. We hope its availability will foster more research interest in studying sustainability in OSS projects.

I. Introduction

Open Source Software (OSS) has become a multi-billion dollar business: 80+% of businesses, including all major tech companies rely on OSS [1]. But OSS projects fail or get abandoned at very high rates, as high as 83% by some accounts, especially the smaller and younger projects [2].

To increase their chances of achieving success, many developers and projects join well-known not-for-profit foundations, like the Apache Software Foundation (ASF). Doing so gets them access to skilled developers and project-specific coaching and mentorship. In return for access to the foundation network and resources, the projects have to abide by the foundation policies and rules, thus give up some of their degrees of freedom. The ASF Incubator (ASFI) is a unique part of ASF, where projects are incubated before being allowed to join into ASF, i.e., before they are deemed successful on their path to self-sustainability. Each project in the incubator has a status, i.e., 'graduated', 'retired', or 'in incubation', corresponding to a successful outcome (i.e., graduation) of joining the ASF, an unsuccessful outcome (i.e., retirement), and still in incubation, respectively. The graduation and retirement status are determined by experienced administrators and contributors in the ASF community, based on how well a project meets a set of goals in terms of sustainability.

Studying success and sustainability in OSS projects is of importance as it can reveal the determinants of successful collaborative work [3], [4]. It also can enhance related social theories which help to understand and reduce issues inherent to building and maintaining diverse technical communities [5]. Rich datasets containing examples of OSS projects that have achieved their goals and those that have not, are essential for such research.

Here we present the ASF Incubator (ASFI) dataset. It contains historical trace data of committer and project activities for 269 ASFI projects that have entered, passed through and exited ASFI. It includes emails, commits, sponsor information, and the incubation outcome (graduated or retired). After reflecting on prior work on sustainability and success of OSS, we give the details of our dataset, scraping methodology, and storage, followed by two potential research studies that can benefit from this data. Our dataset along with the scripts we used to scrape it is available at Zenodo: https://doi.org/10.5281/zenodo.4480753.

II. Prior Work

There has been substantial work on modeling the outcome of OSS projects [6], [7], [8]. A combined socio-technical perspective has shown to be uniquely helpful when analyzing OSS projects [9], [10], specifically for socialization dynamics [11], project quality over time [12], and coordination [13]. That perspective is in line with ASF's key concept of "projects are communities", manifested through the links between people and code. Software engineering researchers often evaluate the outcomes of OSS projects [4] from two angles, the development process perspective, and from the user and developer side, including contributor growth [14], community participation [15], and communication patterns [16]. Many researchers have been interested in the way OSS projects get inducted into the Apache Software Foundation (ASF) [17], [18], [19]. In particular, Duenas et al. have compared the Apache and Eclipse OSS incubators and shown that some governance elements help more than others [20].

III. ASF Incubator Dataset

"If it didn't happen on the mailing list, it didn't happen", is the motto of ASF. In the ASF community, committers are

required to publicly discuss first before making changes to their collaborative artifact (even if they are beneficial). As a unique unit of ASF, the ASF Incubator (ASFI) is a place for nascent projects that want to join ASF. The ASFI, like ASF itself, is built around democratic mechanisms to help projects grow coherent and healthy communities around them. To graduate from ASFI and eventually join ASF, incubator projects are required to show they adopt workflows and governance supportive of a self-sustainable community.

A. Motivation and Originality

The ASF community puts great efforts into keeping and maintaining complete historical development records publicly available. Consequently, ASF developer trace data tends to be at least as reliable as OSS project data from other social coding sites. In addition to the internal reliability of the ASF incubator dataset, ASF committees assign extrinsic binary labels to projects before their incubation exit, graduated (successful), and retired (unsuccessful). This dataset offers heretofore unprecedented opportunities for modeling and comparison of project evolution and sustainability over time.

Moreover, the dataset is longitudinal and includes for each project the most important, nascent stage of development, including the continuous recruitment of new committers, detailed discussions with project mentors, and collaborative actions to address issues. The data runs until the project's incubator exit. This is in contrast to many datasets of GitHub projects which start with well-established infrastructure and previously formed teams and only capture a period of relative stability. Another benefit of studying the ASF incubator projects is that the incubation outcome is evaluated by one consistent and coherent technical community, which can reduce some of the systematic risks of evaluation.

To the best of our knowledge, this is the first time this comprehensive, and well-structured ASFI sustainability dataset has been presented to the empirical software engineering community.

B. Data Source

The Apache mailing list archives can be accessed through the archive Web Page. They contain all the emails and commits from the project's ASFI entry date, and are kept current. The archives are open access, and there do not seem to exist any downloading constraints.

We constructed URLs for individual project files in the ASF incubator as *Project*. The project URLs use the following pattern: project name/(YYYYMM).mbox. For example, for the project *hama*, the full URL is *hama-dev/201904.mbox*. Each such file contains a month of messages in the mailing list of the project, for the date specified in the URL. Here *dev* stands for 'emails among developers'. There are some other common mailing lists, e.g., 'commits', 'issues', 'notifications', 'users' (emails between users and developers or other users). However, many projects, especially those over ten years old which used SVN, use a bot in 'dev' mailing to record all the commits, therefore a message from 'dev' is not always an email from a real person. Similar things were sent to the "commits" mailing list, which, thus, contains some emails. We collected both the *'dev'* and *'commits'* mailing lists files for ASF Incubator projects, through the above archive web page from March 2003 to August 2019.

Online, the ASFI data is organized by calendar month, each month's folder containing the aggregated activities for that month. In some folders there are multiple pages of emails during that month, so we control the pagination by specifying the thread number (e.g., 'thread?0' is the first page, 'thread?1' the second, etc.). The ASF site only displays partial user domains in all email addresses. To get the full email address of the developers (for identifying unique committers), we use the 'raw message' of the emails, which contains the full mbox file.

C. Methodology

We use the well-known *BeautifulSoup* package and the *urllib2* package for collecting and parsing the data from the data sources discussed above, and use the *pg8000* package in *Python* for connecting to a *PostgreSQL* database. Specifically, for commit data, we gathered the following fields: *'From'*: The developer who sent the message, and usually contains both the committer's full name and the associated email address. *'To'*: The mailing list email address of the project. *'Date'*: The exact date (in seconds) when the email was sent. *'Message-ID'*: the unique ID assigned to this email. *'In-Reply-To'*: the message ID that this email replies to. *'Subject'*: The tile of the email. *'Body'*: The body of the email.

Additionally, we gathered commit data from the 'commits' folder, the gathered fields are: *'Author'*: The author of the code. *'Committer'*: The committer who commits the code. The 'Committer' usually is the same as the 'Author'. *'Date'*: the date that the committer pushed the commit. Code Changes: This contains a list of files modified with this commit. It also contains Lines of Code (LOC) changed. A '+' sign at the beginning of a code line represents a new line inserted, while a '-' sign indicates code deletion.

ASF manages and records the communications among people by globally assigning an exclusive project-specific email address to each developer in a project. However, empirical evidence suggests [21] that some developers still prefer to use their personal email/name, which can present ambiguity when identifying distinct developers. To address this issue, we perform de-aliasing for those developers with multiple aliases and/or email addresses. The process is as follows. We first remove titles (e.g., jr.) and common words in the name (e.g., admin, lists, group) from usernames, then we match with both the original order and switched first/last name order whenever names contain exactly one comma (e.g., 'foo, bar' to 'bar, foo'). Then we match each developer with their email address(es) using text similarity.

D. Storage

We store our data in a Postgres database and provide it as CSV tables. The associated data schema is shown in Figure 1.

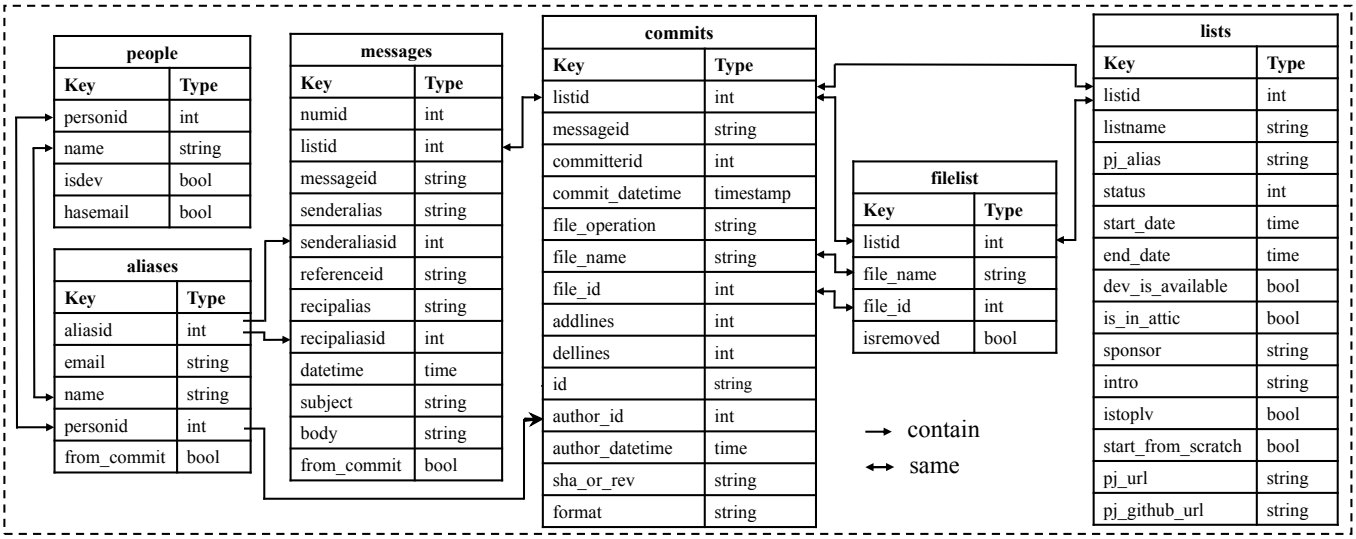

Fig. 1. Data Schema of the ASF Incubator Dataset.

Each table in the schema corresponds to one CSV table in the dataset.

Table ***lists*** contains the project summary from the project homepage. (In the CSV tables, each row represents a project). The keys are as follows. *listid*: project id, a numeric id used as a primary key for indexing, no practical meaning. *listname*: a string, the official project name. *pj_alias*: the project alias used in the archives. *status*: numeric; 0, 1, or 2 corresponding to 'in incubation', 'graduated', and 'retired'. *start_date*: the date that the project entered the incubator. *end_date*: the date that the project leaves the incubator. *dev_is_available*: True if there exists at least one email during the incubation, false otherwise. *is_in_attic*: True if the project had been graduated but was marked as 'reached its end of life'. *sponsor*: the name of the project's sponsor. *intro*: a brief introduction to the project. *istoplv*: A Boolean value, true if the project a top-level ASF project. *start_from_scratch*: A Boolean value, true if the project has started just before entering the incubator. *pj_url*: A string, the homepage URL of the incubator project. *pj_github_url*: A string, the associated GitHub URL for the project. Note that some projects have multiple GitHub repositories and the URLs may change over time.

Table ***messages*** contains the emails ASF committers send, each row represents an email. The keys are listed as follows. *numid*: A numeric id, used as primary key for indexing. *listid*: project id, a numeric id for the project. *messageid*: A string, the message-id, a unique sequence for the email. *senderalias*: A string, the user name of the sender extracted from the email. *senderaliasid*: A numeric id for the sender's alias. *referenceid*: A string, the message-id that this email replies to, blank if it starts a new thread. *recipalias*: A string, the alias of the receiver. *recipaliasid*: A string, the alias id of the receiver. *datetime*: Date timestamp, the date that the email was received. *subject*: A string, the title of the email. *body*: String, the body (content) of the email. *from_commit*: Boolean, true if the email is extracted from commit archives.

Table ***commits*** contains the commits information, each row represents a commit. *listid*: a numeric id for project. The keys are listed as follows. *messageid*: a unique sequence for the message. *committerid*: a numeric id for the committer. A committer is the one who reviews and merges the code to the repository. *commit_datetime*: The date time-stamp of the commit. *file_operation*: A string, including the four operations: (a) new: add a new file to the project; (b) mod: make modifications to an existing file; (c) rm: delete a file from the project; (d) copy: copy a file to another folder without changing any code. *file_name*: A string, the full path to the file. *file_id*: A numeric id for a file, unique in a project. *addlines*: A numeric value, the number of lines of code added. *dellines*: A numeric value, the number of lines of code deleted. *authorid*: A numeric id for the author who writes the code. In the SVN system, the author is the same as the one who commits the code. *author_datetime*: The timestamp of the pull request. *sha_or_rev*: A string, the unique id of the commit, a 'SHA' sequence, or a revision number. *format*: A string, the format of the message in the version control system, 'git' or 'svn'.

Table ***filelist*** contains the histories of file modifications. Note that one file can be deleted and then recreated later. In this case, they are considered as two separate files. In the table, each row represents a file, and the keys are listed as follows. *listid*: A numeric project id for indexing. *file_name*: A string, the full path of the file. *file_id*: A numeric id for a file, unique in a project. *isremoved*: A Boolean value, true if the file has been deleted.

Table ***aliases*** contains developers aliases. Each row in the CVS tables represents an alias. *aliasid*: A numeric id for an alias. *email*: A string, the email address associated with the alias. *name*: A string, the alias. *personid*: A numeric id for the committer who uses this alias. *from_commit*: A boolean, true if the alias is gathered from commits.

TABLE I
THE PROJECT-LEVEL STATITICS OF THE 269 ASF INCUBATOR PROJECTS. THE *incubation_time* (IN MONTH) AND *grad_status* ARE CALCULATED WITHOUT THE PROJECTS THAT ARE STILL IN INCUBATION

Statistic	Mean	St. Dev.	Min	Max
incubation_time	23.93	16.23	1	110
num_commits	1,845.35	3,991.07	0	55,278
num_committers	20.19	24.37	0	161
num_emails	5,395.06	7,943.08	9	94,772
num_senders	176.46	211.42	5	2,104
num_files	7,926.37	13,722.87	0	119,111

E. Metrics

Among the collected 269 projects, 44 are still in incubation, 179 were graduated, and the rest 46 projects were retired. We compute a set of standard software engineering metrics for OSS project activity from the ASFI dataset. These include: incubation time (*incubation_time*): months in the ASF incubator before graduation or retirement); number of commits (*num_commits*); number of committers (*num_committers*); number of emails (*num_emails*); number of email senders (*num_senders*); and number of files (*num_files*). The descriptive statistics for these metrics over the 269 ASF projects are given in Table I[1].

IV. POTENTIAL RESEARCH QUESTIONS

In this section, we present two potential research questions that can be studied based on the ASF incubator dataset.

A. Case I: Studying Sustainability

The ASFI project data can be used to study and understand software engineering metrics that contribute to modeling OSS project sustainability.

In empirical software engineering studies, there have been multiple approaches to define OSS success based on intrinsic project metrics. Some have focused on the technical aspect, e.g., the project development cycle. Others have focused on the social aspect, e.g., community building and popularity. Either direction is justifiable, both from a research and project perspective, as there is no universally agreed definition of OSS success/sustainability.

The dataset presented here contains a different, extrinsic-based sustainability labeling for each OSS project. Upon exit of each incubator project, ASF committee members vote for its graduation and eventually label it as such. Such labeling is done by ASF experts and thus does not have the issues associated with intrinsic, metric-based approaches.

Methodologically, to understand the determinants of sustainability in OSS projects, researchers can use this dataset to regress the extrinsic graduation labels over various project, process, and social metrics of interest in software engineering. Such models, if effective, can enable OSS projects to have a finer level of introspection, over time.

[1]Mins of zero are due to the early retirement of one project (Kabuki)

B. Case II: Studying Effects of Interventions

In ASFI projects, project developers intervene in the direction of the project constantly. For example, a big release can impact projects if everyone takes a break, as delayed maintenance, support, and technical debt become significant when everyone comes back after not working for a while. Project mentors can also affect projects, as they discuss and advocate for the projects. Mentors typically intervene in the incubator projects when projects are inactive for a while by, e.g., asking for status reports. Cursory examinations of commit levels show increases in the periods following a submitted status report. Lastly, when new committers join the projects, they bring in new knowledge and can potentially change the dynamics of the project.

The above-mentioned *intervention events* can presumably be detected from discussions on the mailing lists, which are included in this dataset. Methodologically, this can be done with keyword searches or more sophisticated AI/NLP-based methods. By collecting such longitudinal intervention event data, researchers can study quasi-effects of those interventions on project graduation and suggest potential project-specific recommendations, to aid nascent OSS projects on their road to sustainability.

V. LIMITATIONS AND CONCLUSION

Limitations We note that this dataset has the following limitations: a relatively small scale, of hundreds of projects only; an inherent imbalance between graduated and retired projects; no information on geographical location; and limited diversity (all projects are under similar policies and guidance of the ASF); and there is a potential risk in identifying and merging different developer name (i.e., name disambiguation), especially when a project contains two or more developers sharing the same name.

And lastly, since the incubator projects are all under ASF regulation, generalizing the research implications beyond the ASF community carries potential risks. Thus, further work on expanding the dataset beyond ASF, e.g., with additional OSS projects from GitHub can aid in lowering such risk.

Conclusion Research into OSS project sustainability and success can present actionable insights for maintaining the community. However, there is a dearth of data that are dynamic and have extrinsic sustainability labels upon the exit of project incubation. In this paper, we presented such a longitudinal dataset of technical contributions and developer communication in ASF incubator projects, more narrow in scope than general GitHub projects but with extrinsic, graduation success labels. The presented longitudinal dataset can be used to study why some nascent projects succeeded under the regulation of the ASF incubator community while others do not. Beyond the commits data and email data, in the future, we seek to provide the empirical software engineering community with other OSS sustainability related data, and we will continuously update our dataset when further projects become available.

REFERENCES

[1] L. E. Hecht, "Survey: Open source programs are a best practice among large companies," 2018, [Online; retrieved 31 Jan 2020]. [Online]. Available: https://thenewstack.io/survey-open-source-programs-are-a-best-practice-among-large-companies

[2] C. M. Schweik and R. C. English, *Internet success: a study of open-source software commons*. MIT Press, 2012.

[3] A. Amrollahi, M. Khansari, and A. Manian, "How open source software succeeds? a review of research on success of open source software," 2014.

[4] A. H. Ghapanchi, A. Aurum, and G. Low, "A taxonomy for measuring the success of open source software projects," *First Monday*, vol. 16, no. 8, 2011.

[5] J. Wang, "Survival factors for free open source software projects: A multi-stage perspective," *European Management Journal*, vol. 30, no. 4, pp. 352–371, 2012.

[6] V. Midha and P. Palvia, "Factors affecting the success of open source software," *Journal of Systems and Software*, vol. 85, no. 4, pp. 895–905, 2012.

[7] J. Piggott, "Open source software attributes as success indicators," *Univ. of Twente*, 2013.

[8] C. Subramaniam, R. Sen, and M. L. Nelson, "Determinants of open source software project success: A longitudinal study," *Decision Support Systems*, vol. 46, no. 2, pp. 576–585, 2009.

[9] J. D. Herbsleb, "Global software engineering: The future of socio-technical coordination," in *Future of Software Engineering (FOSE'07)*. IEEE, 2007, pp. 188–198.

[10] W. Scacchi, "Socio-technical interaction networks in free/open source software development processes," in *Software process modeling*. Springer, 2005, pp. 1–27.

[11] N. Ducheneaut, "Socialization in an open source software community: A socio-technical analysis," *Computer Supported Cooperative Work (CSCW)*, vol. 14, no. 4, pp. 323–368, 2005.

[12] C. Bird, N. Nagappan, H. Gall, B. Murphy, and P. Devanbu, "Putting it all together: Using socio-technical networks to predict failures," in *2009 20th International Symposium on Software Reliability Engineering*. IEEE, 2009, pp. 109–119.

[13] C. Bird, "Sociotechnical coordination and collaboration in open source software," in *2011 27th IEEE International Conference on Software Maintenance (ICSM)*. IEEE, 2011, pp. 568–573.

[14] M. S. Zanetti, I. Scholtes, C. J. Tessone, and F. Schweitzer, "The rise and fall of a central contributor: Dynamics of social organization and performance in the gentoo community," in *2013 6th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE)*. IEEE, 2013, pp. 49–56.

[15] N. McDonald and S. Goggins, "Performance and participation in open source software on github," in *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, 2013, pp. 139–144.

[16] J. Wu, K.-Y. Goh, and Q. Tang, "Investigating success of open source software projects: A social network perspective," *ICIS 2007 Proceedings*, p. 105, 2007.

[17] P. C. Rigby and A. E. Hassan, "What can oss mailing lists tell us? a preliminary psychometric text analysis of the apache developer mailing list," in *Fourth International Workshop on Mining Software Repositories (MSR'07: ICSE Workshops 2007)*. IEEE, 2007, pp. 23–23.

[18] M. C. Júnior, M. Mendonça, M. Farias, and P. Henrique, "Oss developers context-specific preferred representational systems: A initial neurolinguistic text analysis of the apache mailing list," in *2010 7th IEEE Working Conference on Mining Software Repositories (MSR 2010)*. IEEE, 2010, pp. 126–129.

[19] M. Gharehyazie, D. Posnett, B. Vasilescu, and V. Filkov, "Developer initiation and social interactions in oss: A case study of the apache software foundation," *Empirical Software Engineering*, vol. 20, no. 5, pp. 1318–1353, 2015.

[20] J. C. Dueñas, F. Cuadrado, M. Santillán, J. L. Ruiz *et al.*, "Apache and eclipse: Comparing open source project incubators," *IEEE software*, vol. 24, no. 6, pp. 90–98, 2007.

[21] J. Zhu and J. Wei, "An empirical study of multiple names and email addresses in oss version control repositories," in *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*. IEEE, 2019, pp. 409–420.