Two Differing Approaches to Survival Analysis of Open Source Python Projects

Derek Robinson
Keanelek Enns
Neha Koulecar
Manish Sihag
drobinson@uvic.ca
keanelekenns@uvic.ca
nehakoulecar@uvic.ca
manishsihag@uvic.ca
University of Victoria
Computer Science
PO Box 1700 STN CSC
Victoria, British Columbia, Canada V8W 2Y2

ABSTRACT

[Keanu: Abstract Pending]

CCS CONCEPTS

• **Software and its engineering** → **Open source model**; • **Information systems** → *Data mining*;

KEYWORDS

data science, survival analysis, open source, python, Kaplan Meier, Cox proportional hazards model, Bayesian analysis

ACM Reference format:
Derek Robinson, Keanelek Enns, Neha Koulecar, and Manish Sihag. 2021. Two Differing Approaches to Survival Analysis of Open Source Python Projects. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* **??** pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The developers of Open Source Software (OSS) projects are often part of decentralized and geographically distributed teams of volunteers. As these developers volunteer their free time to build such OSS projects, they likely want to be confident that the projects they work on will not become inactive. Suppose OSS developers are aware of key attributes that are associated with long-lasting projects. In that case, they can make informed assessments of a given project before devoting their time to it, or they can strive to make their own projects exhibit those attributes. Understanding which attributes of an OSS project lead to its longevity motivated Ali *et al.* to apply survival analysis techniques commonly found in biostatistics to study the probability of survival for popular OSS Python projects [**?**]. Ali *et al.* (referred to as the original authors from here on) specifically studied the effect of the following attributes on the survival of OSS Python projects: publishing major releases, the use of multiple hosting services, the type of hosting service, and the size of the volunteer team.

Survival analysis is a set of methods used to determine how long an entity will live (or the time to a given event of interest, such as death) and is often used in the medical field. For example, survival analysis can determine the probability of a patient surviving past a certain time when given a treatment. However, death is not as well defined for a software project as it is for a living organism. A project may not receive revisions for an extended period only to be returned to at a later date, or perhaps a project no longer receives revisions at all, but the community that uses it continues to be active. Samoladas *et al.* considered a project inactive if it received less than two revisions a month; two months of inactivity led to it being considered abandoned or dead [**?**]. Evangelopoulos *et al.* and the original authors considered a project dead once there were no revisions at all [**? ?**]. The latter definition of project abandonment or death is used in this study to measure the duration of a project or its survival.

The original authors use a frequentist approach to survival analysis utilizing such methods as the Kaplan-Meier (K-M) survival estimator and the Cox Proportional-Hazards model [**? ?**]. Though frequentist approaches are considered to be unbiased, minimal in variance, efficient, and generally sufficient, some consider them to lack robustness [**?**]. Another approach to survival analysis, Bayesian analysis, is considered to generate more robust models that perform well under the presence of new data being introduced and are generally easier to interpret results from [**?**].

The authors of this paper resonate with the motivation of the original authors. This paper serves as a replication of their paper [**?**] (referred to as the original paper from here on) and seeks to assess the validity of their analysis. This replication also provides artifacts so that others may see how the study was conducted and reproduce it with ease. For the sake of the authors' curiosity, an

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

additional attribute of the data was analyzed: the revision frequency of the project. In addition to the replication, this paper analyzes the same data set using a Bayesian approach to survival analysis as outlined in [**?**] and seeks to compare the results of the frequentist and Bayesian approaches in the same domain. Thus, the research questions this paper answers are as follows:

RQ1. How do major releases, the use of multiple hosting services, the type of hosting service, and the size of the volunteer team affect the probability of survival of an OSS Python project?
RQ2. How does the revision frequency of an OSS Python project affect the probability of its survival?
RQ3. How do the findings of frequentist survival analysis differ from Bayesian survival analysis?

Section **??** outlines other research which has utilized survival analysis to study OSS and other work which has studied attributes similar to those studied in this paper. Section **??** describes the source of the data set, the data set itself, and the required preparation in order to perform survival analysis. Section **??** covers the methods used for the replication, additional frequentist analysis, and Bayesian analysis. Section **??** shows the results for each analysis. Section **??** discusses the comparative differences between the analyses performed, the limitations of the study, implications, and suggestions for future work on the topic. The final section concludes by summarizing the purpose and findings of this paper.

2 RELATED WORK

Several other researchers have employed survival analysis to study the health of OSS projects. For example, Samoladas *et al.* studied the affect of application domain and developer count on OSS project health [**?**]. They found that applications within the domain of *games and entertainment* and *security* had the lowest probability of survival. Additionally, they found that for each new developer introduced to a project, the projects survivability increased by 15.8%. On the topic of developers, several studies have used survival analysis to study developer disengagement from OSS projects [**? ? ?**]. Miller *et al.* made use of a survey and survival analyses, to determine the causes behind why a developer might stop contributing to an OSS project [**?**]. Their analysis revealed that developers have a higher probability of project disengagement when going through job transitions and when working longer hours. Lin *et al.* determined that developers who balance maintaining files they have created with maintaining files others have created have a higher survival probability than developers who only maintain their own files or only maintain others files [**?**]. Additionally, Lin *et al.* found that developers who maintained files and developers who mainly wrote code had a higher survival probability than those who solely created files and those whose main focus was writing documentation. Ortega and Izquierdo-Cortazar analyzed the survival of FLOSS committers and Wikipedia editors and found that FLOSS committers have higher mean survival times than Wikipedia editors [**?**]. Survival analysis can also be applied to the software it self, this has been demonstrated by Aman *et al.* and Caivano *et al.* [**? ?**]. In [**?**], survival analysis was used to analyze time to bug-fix for files modified by developers of different experience levels. This analysis determined that files which most recently modified by less experienced developers had an increased probability of needing a bug fix within a shorter time frame [**?**]. Caivano *et al.* explored the affect of dead code within OSS projects using survival analysis [**?**]. They found that dead methods are present in Java code, persist for a long time (in terms of commits) before being buried or revived, are rarely revived, and that most dead methods have been dead since their inception.

Other studies have examined the health of OSS projects using methods other than survival analysis. Xia *et al.* predicted a number of health indicators of OSS projects, such as, the number of developers and the number of revisions. These predictions were made using regression trees that were optimized using differential evolution, leading to a 10% increase in prediction accuracy over the base line [**?**]. Norick *et al.* analyzed OSS projects using code quality measures and observed no significant evidence that the number of committing developers affects software quality [**?**].

3 DATA

Performing survival analysis of OSS projects requires a data set that records the repositories for projects on common hosting sites, including a history of all commits (revisions from here on out) and major releases (revisions of note, often with a specific name and release date) [**?**]. The *popular-3k-python* subset of the Software Heritage graph [**?**] contains the necessary information and was used in the original paper and also in this paper. This data set contains information on 3052 popular Python projects hosted on GitHub/GitLab, Debian, and PyPI, and records revisions between 1980 and 2019 at the time of writing (the Software Heritage graph is subject to updates, which makes it a non-reproducible data set). Following the tutorial provided by the Software Heritage organization [**?**], a PostgreSQL database was hosted on the authors' local machines to facilitate data collection.

Though the *popular-3k-python* data set contains all the necessary information to perform survival analysis, it first must be manipulated into a more suitable format before the analysis can be carried out. In this case, the collected data was manipulated such that the final data set contained the duration of the project, the censorship value, and the attributes of interest. Descriptions of each column present in the data set can be found in Table **??**. Data collection and manipulation were performed in Jupyter Notebooks, which are available in this paper's repository, a link to which can be found in appendix **??**.

Table 1: Data Set Column Descriptions

Column Name	Description
Host Type	Which hosting service the project's repository resides o
Major Releases	Whether or not the project publishes major releases
Censored	True if the project's death is not observed (for more info
Duration_months	The duration of the project in months
High_rev_frequency	Whether or not the project has high revision frequency
Multi_repo	Whether or not the project is hosted on multiple hostin
High_author_count	Whether or not the project has a high author count (gre

For their study, the original authors set a time frame of 165 months (where a month is defined as 28 days), starting in 2005 and ending in January 2018. This paper uses the same time frame and

determines exact start and end dates. Using January 1, 2018, as a strict end date and maintaining the study duration as 4620 days (165 months as defined), the start date is found to be May 9, 2005. After following the same procedures described in the original paper, a list of 2066 projects and their associated information was obtained.

4 METHODS

4.1 Replication

4.1.1 Death and Censoring. Two critical concepts in survival analysis are events of interest and censoring. As previously discussed, the event of interest for this study is project abandonment or death. As defined in the introduction, a project is considered dead when it no longer receives any revisions. With this definition, it is impossible to know whether any project is truly dead, because, unlike a living organism, any project may receive revisions at any point in the future, making it "not dead" by the working definition. However, there are multiple practical ways of determining whether a project is dead, all of which rely on the scope of the studied data set. For this study, the death of a project is determined by first defining two special revisions for each project: *last revision* and *last observed revision*. *Last revision* is defined as the last recorded revision of a project within the scope of the data set (i.e. 1980 - 2019). *Last observed revision* is defined as the last recorded revision of a project within the studied time frame (i.e. 2005 - 2018). If, and only if, the *last revision* is also the *last observed revision*, the project death is said to be observed.

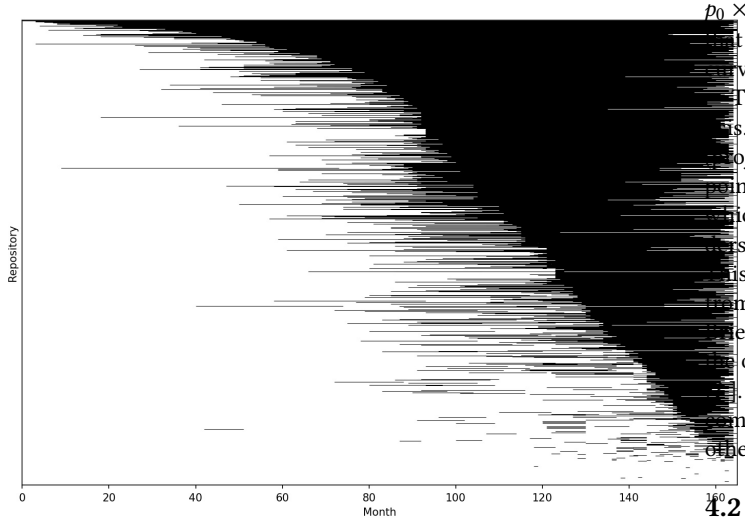


Figure 1: Graph of project durations within the studied time frame. The projects are ordered by duration and plotted from and to their respective start and end dates. Month 0 begins on May 9, 2005, and Month 165 concludes on January 1, 2018. The black portion of the horizontal lines indicates the active period of a given project

What happens if a project's death is not observed? This is where censoring is used. Suppose a project has its first revision one month before the end of the studied time frame, and continues to be revised daily past the studied time frame. It would be incorrect to indicate that this project survived for only one month, but it is also impossible to include the project's future activity in the study. Rather than discarding such projects, censoring causes them to be considered by the study for the observed duration, but without considering them as dead projects (i.e. they are removed from the calculations after they stop being observed). There are multiple types of censoring in survival analysis, but this study uses random censoring or type III censoring, which involves removing subjects from a study at varying times relative to when they began to be observed (as is the case here) [**?**]. If a project's death is not observed, it is considered censored.

The distribution of the project durations over the studied time frame can be seen in Fig. **??**, which is a more detailed rendition of Figure 1 in the original paper. Note that when the black lines extend to the end of the time frame, this likely indicates the project was censored. About 62% of the studied projects were censored.

4.1.2 Survival Analysis. Using both the calculated duration associated with each project and the censoring status of each project, the survival analysis can be performed. The analysis was carried out using an R notebook (this can be found in the project repository under Analysis&Data/frequentist-analysis.rmd).

The K-M estimator is a non-parametric estimation technique that estimates the survival function, $S(t)$. The survival function gives the probability that a given project will survive past a particular time t. At $t = 0$, the K-M estimator is 1 and as t approaches infinity, so does the K-M estimator. More precisely, $S(t)$ is given by $S(t) = p_0 \times p_1 \times p_2 \times \cdots \times p_t$, where p_i is the proportion of all projects that survived at the i^{th} time step [**?**]. The K-M estimator produces curves that approach the true survival function of the data.

The hazard function is another useful function in survival analysis. It describes the probability of the event of interest or hazard (project abandonment in this case) occurring up to a given time point. The original paper uses the Cox Proportional-Hazards model which allows for fitting a regression model in order to better understand how the health of projects relate to their key attributes. This analysis results in the hazards ratio (HR), which is derived from the model for all covariates that are included in the formula. Briefly, a $HR > 1$ indicates an increased risk of abandonment; on the other hand, $HR < 1$, indicates a decreased risk of abandonment [**?**]. As such, the HR represents a relative risk of abandonment that compares one instance of a binary feature (e.g. yes or no) to the other instance.

4.2 Revision Frequency Analysis

The original paper mentions that "The health of a project could be computed by the number and frequency of contributions..." but never addresses this measurement. This paper seeks to explore the frequency of contributions as a method of assessment (simply analyzing the number of contributions would not yield useful results given the varying nature of the project durations in the studied time frame). The revision frequency, defined as the number of commits divided by the number of days in the project's observed lifetime, was dichotomized into two groups depending on whether the frequency was above one revision per day. Although the median revision frequency was approximately 0.68 revisions per day, the threshold value of one was chosen because it is easier to remember

when keeping these attributes in mind and, similar to the other dichotomizing attributes, it provides a threshold that fewer projects attain to, which sets them apart. This study applies both the K-M estimator and the Cox Proportional-Hazards model to the data to stratify the effects of high revision frequency on the overall health of an open-source project.

4.3 Bayesian Survival Analysis

The Bayesian approach to survival analysis is less common due to computational difficulties. However, it offers multiple advantages over the frequentist approach [**?**]. We replicate the methods outlined in [**?**] to apply Bayesian survival analysis to our study. The Bayesian analysis uses posterior distributions of model parameters to draw inferences about them. These posterior distributions are obtained via Markov-Chain-Monte-Carlo (MCMC) algorithms. The statistical modelling language used was Stan.

Our study applies the same methods as found in the section titled *A detailed example* of [**?**]. A parametric exponential model that assumes the survival times of a project $y = (y_1, y_2, \ldots, y_n)$ are exponentially distributed with parameter λ was created. The censoring indicators as $v = (v_1, v_2, \ldots, v_n)$ where $v_i = 0$ if y_i is right censored (project death is not observed) and $v_i = 1$ if y_i is a failure time (project death is observed), the survival function which is the probability of surviving past the time point y_i is given by

$$S(y_i|\lambda) = P(T \geq yi|T \geq 0) = 1 - [1 - \exp(-\lambda y_i)] = \exp(-\lambda y_i) \quad (1)$$

and the survival model is denoted as

$$y_i|v_i \sim f(y_i|\lambda)^{v_i} + S(y_i|\lambda)^{1-v_i} = [\lambda exp(-\lambda y_i)]^{v_i} + [exp(-\lambda y_i)]^{1-v_i} \quad (2)$$

$$\lambda \sim p(\lambda) \quad (3)$$

$$\lambda = exp(x_i^T \beta) \quad (4)$$

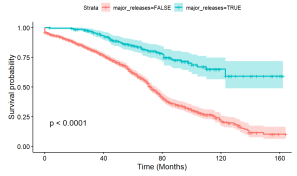
This model was then used to visualize the posterior survival functions for the following five project attributes: major releases, hosting service of the project, use of multiple hosting services, team size, and revision frequency.

5 RESULTS

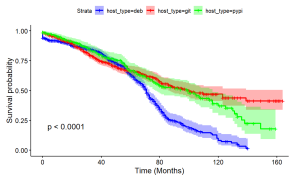
5.1 Replication

The replication study performed in this paper yielded extremely similar results to those shown in the original paper. Fig. **??** depicts K-M curves along with their confidence intervals and p-values. The p-values imply that the difference in survival probability for projects within each group is statistically significant. As seen in Fig. **??**, this study found out that projects with at least one major release have higher chances of survival. The curve for projects with at least one major releases plateaus around 65% survival probability around the 120-month mark whereas the survival probability of projects with no major releases ends up to be less than 20% by the end of the study period. Fig. **??** represents the significance of hosting the project on different hosting services, it was observed that projects that are hosted on GitHub have higher survival chances in the long run, as the curves suggest all three hosting services have a similar trend for the first 55 months which is within the average duration of projects hosted on these services. In addition, having multiple repositories hosted on multiple services has significantly increased the chances of a project's survival. As seen in Fig. **??**, the survival rate for such projects is 75%, whereas around 20% for projects with only one package repository system. The Curve in fig. **??** implies the significance of network of developers for the survival of open source project, the project having more than 20 different authors end up having around 60% survival rate which is significantly different from projects with a small team of developers with less than 20% survival rate at the end of the study period.

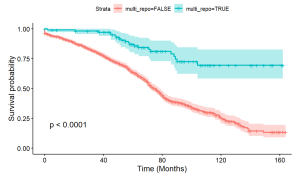
Figure **??** quantify estimates of these attributes' effect on the survival probability using the Cox Proportional-hazards model. The third column shows the hazards ratio which indicates the probability of abandonment with respect to the reference feature. In the first row, the hazard ratio for projects not having a major release with respect to projects having major releases is nearly 3, it implies that the projects without major releases are three times more likely to become inactive compared to projects with at least one major release. Similarly, projects having only one repository system are 3.3 times more likely to be abandoned. The third row highlights that the projects with fewer developers count are 19.3 times more likely to be inactive. For the type of hosting used, the ratio implies that projects hosted on PyPi or Debian are less likely to be abandoned compared to projects that are hosted on GitHub. This appears to contradict the results of the K-M curve and will be further discussed in Section **??**.



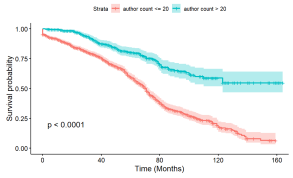
(a) K-M curves for projects which publish major releases and those which do not



(b) K-M curves for projects with different hosting services



(c) K-M curves for projects with repositories on multiple hosting services



(d) K-M curves for projects with high and low author count

Figure 2: K-M curves for each project attribute

5.2 Revision Frequency Analysis

The K-M curves for revision frequency are little different from the graphs we generated for other attributes, as seen in Fig 4. We observed curves drop for the projects with more than one revision per day in the starting of study period. In Addition, the projects with less revision frequency out performed projects with higher

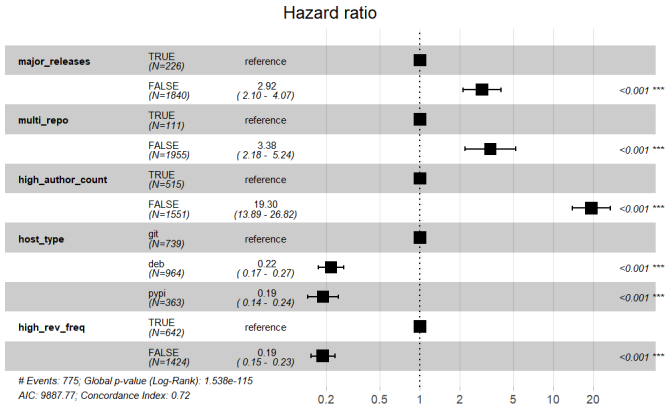


Figure 3: The Cox Proportional Hazards model. From left to right: project attribute, attribute value with counts, hazard ratio, box plot of hazard ratio, and p-value of log rank test.

revision frequencies during the mean duration period of the studied projects. As shown in the Fig. 3, results of our regression model also indicate that the projects with less revision frequency are less likely to be abandoned.

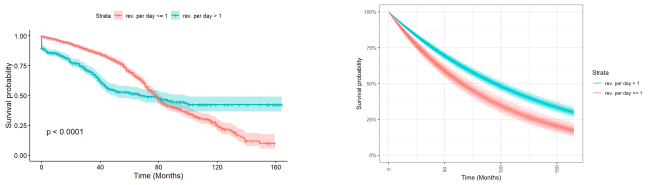
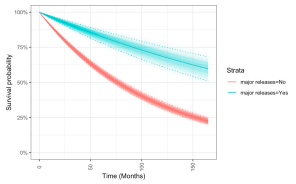


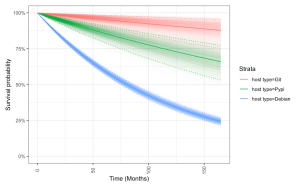
Figure 4: K-M (left) and Bayesian posterior survival function (right) curves for revision frequency

5.3 Bayesian Survival Analysis

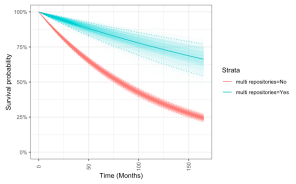
Shown in Fig. **??** are the posterior survival functions for variations of the selected project attributes. The dotted lines represent the 2.5% and 97.5% quantiles, while the solid middle line represents the posterior mean of β (the prior on the project attribute of interest). The remaining lines all represent all valid posterior survival functions. Fig. **??** clearly indicates that the survival function of the projects with no major releases decreases much faster than projects with releases. Survival probability for projects with no major releases was lower than 25% after 150 months compared to 55% for projects with major releases. Fig. **??** illustrates that projects hosted on Github have the highest survival chances compared to those hosted on PyPi and Debian. At 150 months, the predicted survival probabilities for Github, PyPi and Debian were 87%, 65% and 25%, respectively. Fig. **??** shows that projects with repositories on more than 1 hosting service had significantly higher survival chances than projects with repositories on a single hosting service. Above 65% survival probability for multiple repositories compared to slightly over 25% for single repository projects at the end of 150 months. Fig. **??** demonstrates the importance of the number of contributing developers on the survival chances of open-source python projects. For projects with more than 20 authors, the predicted survival probability was around 55%, while less than 25% for projects with less than 20 authors at the 150-month mark. Additionally, for all project attributes, the beta value for the 97.5% quantile was found smaller than zero, which ensures that estimates are highly certain [?]. **[Derek: This last sentence needs a litte work. Some elaboration on what is meant by highly certain should suffice.]**



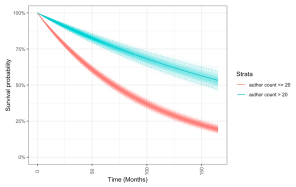
(a) Posterior survival functions for projects which publish major releases and those which do not



(b) Posterior survival functions for projects with different hosting services



(c) Posterior survival functions for projects with repositories on multiple hosting services



(d) Posterior survival functions for projects with high and low author count

Figure 5: Bayesian posterior survival functions for each project attribute

6 DISCUSSION

6.1 Implications

6.2 Comparison of Analyses

[Keanu: We are not quite ready for a comparison yet]

6.3 Limitations

6.3.1 Limitations of the Methods. The original paper [**?**] and the MSR presentation given [**?**] have contradicting methods of censoring. The method discussed in the original paper was deemed superior and is used in this paper.

The original authors left out many details regarding the methods they used for data extraction, manipulation, and analysis. This led to assumptions being made when replicating their study. Hence, there are discrepancies in the values obtained compared to the original paper.

Being a replication, this study was limited to choosing the methods used by the original paper. These methods are also unanimous in the area of survival analysis, and there are few alternatives to them. This study attempts to address these limitations by providing an alternative analysis method, namely Bayesian analysis.

When applying the K-M estimator, it is common to use a log-rank test to test the significance between the two groups which are being compared. The log-rank test only indicates whether or not the probability of survival is statistically significant between the two groups and is not able to provide any information about the size of the difference between the two groups [**?**]. Additionally, the K-M estimator does not account for confounding factors [**?**]. In more traditional uses of the K-M estimator, an example of a confounding factor could be the age of the study participants. In the case of this study, there may be confounding factors such as the experience level of the developers or whether the developers received funding to work on the project. Neither of these factors are represented in the data set.

The Cox Proportional-Hazards model is used with the assumption that, over the period of observation, the hazards within each group are proportional [**?**]. If the assumption that the hazards within each group are proportional is not true, then the Cox Proportional-Hazards model will lead to incorrect estimates of the hazard ratio between two groups [**?**]. Looking at the K-M curves for this study, it can be concluded that the proportional hazards assumption does not hold as the survival functions diverge over time and cross over each other rather than running in parallel [**?**]. This explains the discrepancies between the results of the cox regression model and the K-M curves. Future studies should perform tests to determine whether the assumptions for their models hold and should seek methods for mitigating such errors through identifying time-dependant covariates.

The Bayesian approach to survival analysis comes with its own limitations as well. As pointed out by Renganathan, Bayesian survival analysis can be subjective as the analyst places their own bias into the model when selecting the prior distributions [**?**]. In order to mitigate this bias, prior selection requires both epistemological and ontological reasoning. Prior distributions were chosen based on survival analysis done in other domains [**? ?**], but more investigation should be done as to whether these priors are appropriate for this domain and whether the models from this study accurately predict durations of different data sets.

6.3.2 Limitations of the Data. The data set in this study has been aggregated from multiple version control systems across the web over a large period of time. As such, the data set is not fully reproducible, as pointed out by the original authors of the Software Heritage organization [**?**]. Additionally, it cannot be ensured that the data contains a full history of the respective repositories. The lack of certainty about the full history is because the repository admin can modify the history of revisions to suit their liking [**?**].

There are inherent differences in the ways developers use the different hosting services. Traditionally, services such as PyPi and Debian are used to host major releases of a product. This may hide information about the number of developers and the revision frequency. Additionally, the potential confounding factors mentioned in section **??** are not represented in this data set.

It may also be worth noting that this data is only for Python projects and it is possible that different behaviours are associated with development in different languages. Python is a relatively easy language to use, and can often be used for small tasks that are not maintained. The results of the revision frequency analysis in section **??** seem to indicate a large number of short lived projects.

The data set contains a large portion of censored data. This means the abandonment of most of the projects was not observed. As data points are censored (denoted by the vertical tick marks in the KM curves), there is a smaller and smaller group of data points to study. This means that the results towards the 165 month mark may be less representative.

The data set contained many revisions (over 4 million) that were not associated with project URLs. The cause of this is unclear.

6.4 Future Work

[Keanu: Just a couple of ideas so far]

Increase the time frame of the study (the original paper could not do this because the paper was written in 2018).

Perform separate studies on each hosting service to remove variability in the way the services are utilized.

[Derek: Future work should include a more in depth analysis of whether or not the assumption of the cox proportional hazards model holds]

[Keanu: Remove short lived projects, e.g. a day long]

[Keanu: this study considered the first observed revision to be the first revision of a project (similar definitions when compared to those in section ??), but this may not always be the case. Future studies should acknowledge the true beginning revision for projects in order to obtain a more accurate duration value as in [?].]

7 CONCLUSION

APPENDICES

A ARTIFACTS

Project Repository: https://github.com/DerekRobin/CSC578B-Project
Data Set: https://annex.softwareheritage.org/public/dataset/graph/latest/popular-3k-python/sql/

B WORK DISTRIBUTION