Proceedings of the 2021 Winter Simulation Conference
S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo and M. Loper, eds.

CHARACTERIZING THE DISTRIBUTIONS OF COMMITS IN LARGE SOURCE CODE REPOSITORIES

Aradhana Soni

Department of Industrial & Systems Engineering
University of Tennessee
Knoxville, Tennessee 37996, USA

ABSTRACT

Modern software development is based on software repositories and changes committed to those repositories. However, there is an inadequate insight into the nature of changes committed to repositories of different sizes. A data-based characterization of commit activity in large software hubs contributes to a better understanding of software development and can feed into early detection of bugs at the earliest phases (Alali, Kagdi, and Maletic 2008). Here, we present preliminary results from characterizing the distribution of 452 million commits in a metadata listing from GitHub repositories. Based on multiple distributions, we find the best fits and second best fits across different ranges in the data. The characterization is aimed at synthetic repository generation suitable for use in simulation and machine learning.

1 INTRODUCTION

Over the last three decades, there has been a massive adoption of free and open source software including GitHub. This has made source code and development history of millions of software projects available to public (Raymond and O'Reilly 1999). We aim to characterize the 452 million commits metadata of these publicly accessible software projects to get closed-formed best fitting distributions to generate synthetic equivalence suitable for use in simulation and machine learning. Although this is a first step by which we are getting started with this data set, the presented approach is applicable to other data sets as well.

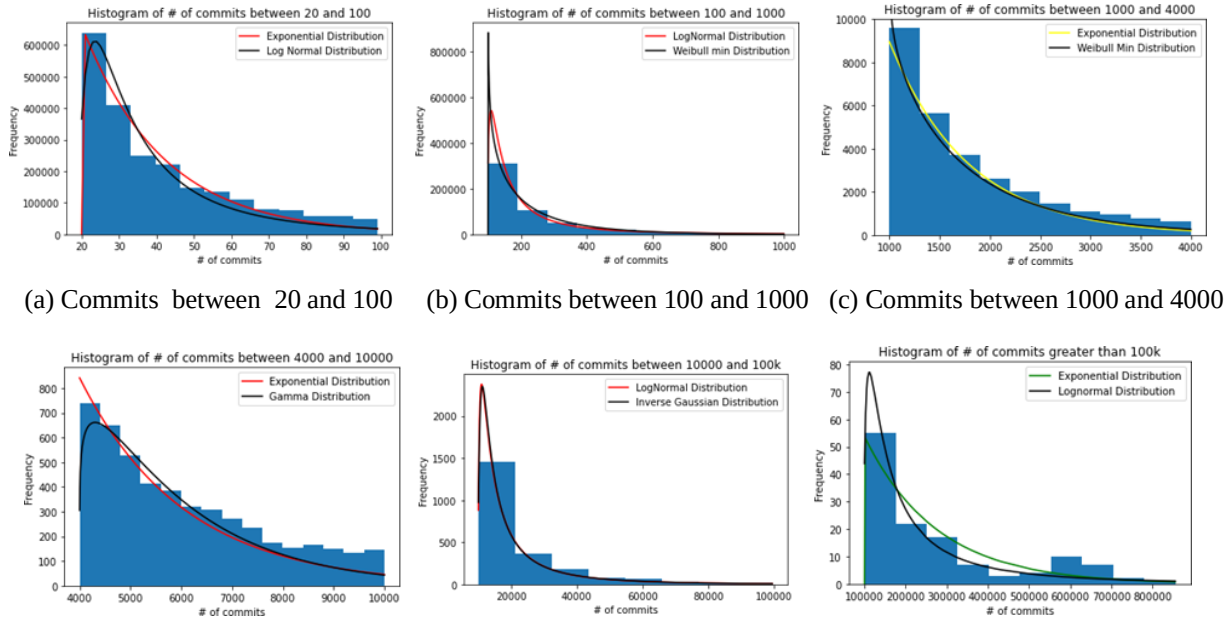
2 DATA SOURCES AND PROCESSING

Our research uses the metadata of 452 million commits to 16 million GitHub repositories to study the commits and contributors. The repositories have been grouped by the number of commits to study the distributions of repositories according to the level of activities in those repositories. The groups range from repositories with fewer than 20 commits to others with over 100 thousand commits. Figure 1 displays the histograms of the number of commits in each group. We have used Python's Scipy Library to identify and approximate the best fitting distributions of "number of commits" in each of these groups. We used Chi-square statistics to check for the Goodness of fit and rank the distributions by their fit.

3 DISTRIBUTIONS AND NEXT STEPS

Table 1 provides the list of distributions that best fits each group along with the parameters. We provide location and scale parameters for exponential distribution and location, scale and shape parameters for all other distributions. For most of the groups, exponential and/or lognormal appears to be the best fitting distribution. Distributions for some groups appear to fit the data better than the others. We understand that this real-world data may not follow any well known probability distributions exactly. Hence we approximated the most probable probability distribution for each of the group and checked it's Goodness of fit with the help of Chi-squared statistic and ranked the distributions on the basis of this statistic. For the

next steps, we will feed these distributions into a simulation of commit activity for synthetic data generation. This data will feed as training data into a machine learning system based on graph neural networks (Hu et al. 2020), which ultimately is aimed at intelligent, real-time alerts and tagging of commits with respect to their susceptibility for errors, bugs, etc.



(a) Commits between 20 and 100 (b) Commits between 100 and 1000 (c) Commits between 1000 and 4000

(d) Commits between 4000 and 10000 (e) Commits between 10000 and 100000 (f) Commits more than 100000

Figure 1: Histograms of Commits.

Table 1: Distributions of commits by number of repositories.

# of commits	# of Repositories	Best Fit	Second Best Fit
<20	13,156,036	Exponential (-0.83,0.83)	Lognormal (5.67,-0.832,0)
20 - 100	2,235,831	Exponential (-1.07,1.07)	Lognormal (1.01,-1.17,0.76)
100 - 1000	554,079	Lognormal (1.30,-0.83,0.41)	Weibull Min (0.81,-0.81,0.71)
1000 - 4000	28,549	Exponential (-1.07,1.07)	Weibull Min (0.93,-1.07,1.11)
4000 - 10,000	4,766	Exponential (-1.26,1.26)	Gamma (1.17,-1.26,1.07)
10,000 - 100,000	2,221	Lognormal (1.30,-0.81,0.40)	Inverse Gaussian (2.13,-0.851,0.40)
>100,000	128	Exponential (-0.94,0.94)	Lognormal (1.33,-0.96,0.48)

ACKNOWLEDGEMENTS

This research was performed under the sponsorship and advisement of Prof. Kalyan Perumalla, Dept. of Industrial and Systems Engg., University of Tennessee, Knoxville, and in collaboration with Rupam Dey, a doctoral student in the same department.

REFERENCES

Alali, A., H. Kagdi, and J. I. Maletic. 2008. "What's a typical commit? A characterization of open source software repositories". In *the 16th IEEE international conference on program comprehension*, June 10th-13th, Amsterdam, Netherlands, 182-191.

Hu, W. et al. 2020. "Open graph benchmark: Datasets for machine learning on graphs". *arXiv preprint arXiv:2005.00687.*

Raymond, E. S., and T. O'Reilly. 1999. *The Cathedral and the Bazaar*. 1st ed. USA: O'Reilly amp; Associates, Inc.