ABSTRACT

Mining GitHub to Identify Open-Source Software Health in Blockchain Projects

Jeff Njisse

Keywords: open-source software, GitHub, blockchain, structural equation modelling

GitHub represents the largest open-source software (OSS) community hosting over 100 million repositories by 56 million developers. Since the rise of Bitcoin [1], blockchain projects have grown dramatically and 84% of the top 200 are hosted on GitHub. Growth inspires many new initiatives and clones, but few projects survive. Identifying healthy projects can assist developers looking to contribute to OSS and researchers seeking innovative teams in addition to reducing fraudulent activity. This presentation is part of a larger ongoing study investigating OSS blockchain performance and health. The hypothesis to be studied is that the health of an OSS project can be determined through publicly available data. This will be investigated by answering the research question: How can factors that influence the health of OSS be identified? The methodology is adapted from a framework for extracting, processing, and analysing OSS [2] that combines eleven variables into a measurement model. Latent variables are introduced to represent the constructs of community engagement, project robustness, and public interest. Statistical analysis methods are employed both for the development and specification of the measurement model and the construction and evaluation of its structural efficacy. Validation is by two mechanisms: confirmatory factor analysis applied to the measurement model and structural equation modelling to estimate the validity of the relationships [3]. In the broader study a tool is to be developed to monitor health and highlight areas for innovation. Preliminary findings indicate that community engagement and project interest are positively correlated to robustness, which, in turn, is expected to be a predictor of software project health. A future hypothesis is that healthy projects can be a leading indicator of innovative technology and practice. If OSS communities and projects can be monitored for health, stake holders can identify areas of strength and weakness, extending to industries beyond blockchain.

References

[1]   S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," 2008. https://bitcoin.org/bitcoin.pdf

[2]   M. Goeminne and T. Mens, "Analyzing ecosystems for open source software developer communities," in Software Ecosystems: Analyzing and Managing Business Networks in the Software Industry, no. 2013, S. Jansen, Ed. 2013, pp. 247–275.

[3]   J. F. Hair Jr., W. C. Black, B. J. Babin, and R. E. Anderson, Multivariate Data Analysis, Seventh. Essex: Pearson Education Limited, 2014.