Diversity in Github: A Data Visualization Study*

Karina Kohl Silveira
School of Technology, PUCRS
Porto Alegre, Brazil
karina.kohl@acad.pucrs.br

César Souza Liedke
School of Technology, PUCRS
Porto Alegre, Brazil
cesar.liedke@acad.pucrs.br

ABSTRACT

A lot has been said about the importance of talking about diversity in Computer Science and Software Engineering. There is a clear lack of representativeness when we observe gender distribution in Information Technology jobs and students in universities, for example. Diversity is good beyond ethic reasons, it's recognized as valuable and, lot of studies have been done about it. Large technology companies have been creating annual reports of their efforts to have a more diverse workforce, increasing minority numbers through recruiting, working to minimize unconscious bias and also investing in programs to increase representativeness. In this paper, we present a work in progress study that links a definition of diversity to the results of a survey applied to explore the perceptions of diversity in agile software development teams in Brazil and an initial analysis focusing mainly on the role of women in those teams.

KEYWORDS

Visualization, GitHub, Diversity

ACM Reference Format:
Karina Kohl Silveira and César Souza Liedke. 2018. Diversity in Github: A Data Visualization Study. In *Proceedings of Visualização de Dados (VD'18).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

More than ever, there are lots of information available that allow us to work in powerful data visualization and GitHub is a rich field to get great datasets.

Vasilescu et al [7] presented for the first time a large data set of social diversity attributes of programmers in GitHub teams. They collected a team social diversity dataset of 23,493 GitHub projects and used alias resolution, location data, and gender inference techniques. They illustrated how the data set could be used in practice and the utility for studying the relationship between social diversity and technical activity in online teams with a series of case studies. In their paper, they provided some stacked bars to exemplify the data set usage, but there is space to enrich the research using different visualization techniques to better understand the data and the suggested case studies. With this motivation in mind, we choose two case studies suggested in their paper to provide new ways to visualize and interact with data and we will better explain it in Section 3.

2 RELATED WORK

Some other works use GitHub data to provide visualization methods to solve some business problems. In [2], they provide ways to visualize GitHub developers data to support the process of recruit appropriate developers for software projects once traditional resume only shows experiences claimed by developers, and very few evidence or information regarding their actual development activities can be obtained. They propose an approach to extract developers' practical activities from their participated open source software(OSS) projects and generate the biographies that reflect their OSS contributions.

Some other kinds of data from social interactions are also largely used as data sets. In [1] they use Reddit.com data to model process that connects the demographic groups with features that best describe the distinguishing characteristics of each group.

3 VISUALIZATION TECHNIQUES AND DATA INTERACTION

The Diversity dataset [6] used in this work contains data from 23,493 projects and a total of 93,056 rows of content on the composition, characteristics, and outcomes of their teams of contributors from 2009 to 2013. The data is available in CSV format.

We combined different visualization techniques to correlate the information of the data set and provide insights from them. For the Case Study mentioned in 3.1 we implement an overlay bar graph with two variables for every project. For Case Study mentioned in 3.2 we implement a line graph with two variables separating by language and projects and also the mean from all projects from that language, using the 'All' option.

To implement the visualization we used Plotly for Python [5] and Jupyter [3]. Plotly for Python is a graphing library that supports to make interactive graphs as line plots, scatter plots, area charts, bar charts, error bars, box plots, histograms, heatmaps, subplots, multiple-axes, polar charts, and bubble charts. Jupyter is an open-source web application that allows to create and share documents that contain live code, equations, visualizations and narrative text. Some uses for Jupyter are data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, etc.

3.1 Case Study 1: Diversity Across Projects

Accordingly to Vasilescu et al. [7], diversity in a team can be identified when team members are different from each other with respect

*Produces the permission block, and copyright information

VD'18, Jun, 2018, Porto Alegre, BR

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

to some attribute e.g., gender, tenure, nationality, etc. We can use aggregate measures of diversity, such as the Blau index to capture how diverse groups are, i.e., the higher the measures, the more diverse team members are with respect to a given attribute.

3.1.1 Gender and Country. A team consisting only of same gender members cannot consider gender diverse; however, a team can reach its maximal gender diversity by having equally many females and male members (assuming a simplified, binary gender), regardless of team size. Using the Blau index, gender uniformity is encoded as 0 and maximal gender diversity is encoded as 0.5 (since the dataset considers only two possible values, male and female). In Figure 1 we present a Stacked Bar Graph for visualization of the data.

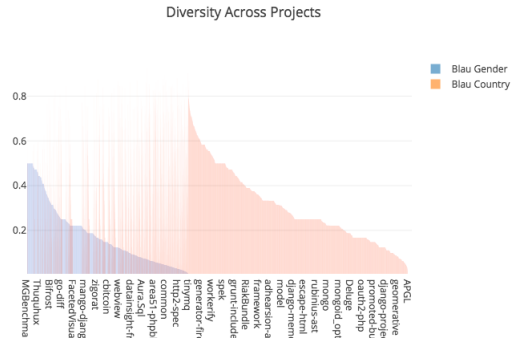

Figure 1: Blau Gender and Blau Country Per Project

3.2 Case Study 2: Diversity Over Time

In this case study we propose the plot of evolutionary setting as the evolution of the Blau index of gender diversity in a project during six years (24 quarters) of history. We can observe how the project's team of contributors was gender balanced in the first two years (Blau index values above 0.4). However, with time, as the contributor team grew (team size is shown dotted for comparison), it also became male-dominated. We present in Figure 2 a Dual Lines Graph for visualization of the data.

4 DISCUSSION

Accordingly to Mackinlay [4] graphic presentations should follow two criteria: expressiveness and effectiveness. With expressiveness we identify the graphic presentations that express the desired information; we prioritize important information and avoid false inferences and we do consistent visual mappings. With effectiveness criteria, we can determine graphic presentations that, in a specific situation, allow us a faster and precise data interpretation. It minimizes the cognitive effort and highlights the main elements.

With the visualizations proposed in this paper, we can faster understand the information under the Diversity dataset we used. From the first Case Study 3.1 we can easily infer which projects were the most gender diverse over the entire period of time considered. Also, with the interaction techniques provided, we can choose specific samples from the less to the most diverse projects in terms

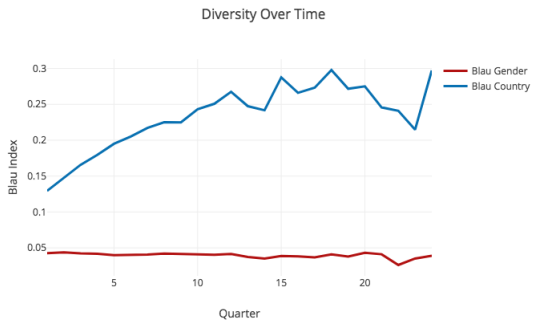

Figure 2: Blau Gender and Blau Country over Time

of gender and country diversity. For the second Case Study 3.2 we can select a very specific period of time and check, in that period, how gender and country diversity changes over the time independent from projects.

5 CONCLUSIONS

In this paper, we propose two visualization techniques to the diversity dataset and two cases studies proposed by Vasilescu et al [7]. We implemented a web-based visualization using Plotly for Python Library, Jupyter.

Clearly, the graphic presentation of the data help us to understand the information and it complements the textual evaluation of the information contained in the data. We used the dataset partially and based on the original paper and the data available, it is possible to provide other correlations and visualizations as well to complement.

REFERENCES

[1] W. Dou, I. Cho, O. ElTayeby, J. Choo, X. Wang, and W. Ribarsky. 2015. DemographicVis: Analyzing demographic information based on user generated content. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 57–64. https://doi.org/10.1109/VAST.2015.7347631

[2] T. Jaruchotrattanasakul, X. Yang, E. Makihara, K. Fujiwara, and H. Iida. 2016. Open Source Resume (OSR): A Visualization Tool for Presenting OSS Biographies of Developers. In *2016 7th International Workshop on Empirical Software Engineering in Practice (IWESEP)*. 57–62. https://doi.org/10.1109/IWESEP.2016.17

[3] Jupyter. [n. d.]. Jupyter. ([n. d.]). Retrieved Jun 21, 2018 from http://jupyter.org/

[4] Jock Mackinlay. 1986. Automating the Design of Graphical Presentations of Relational Information. *ACM Trans. Graph.* 5, 2 (April 1986), 110–141. https://doi.org/10.1145/22949.22950

[5] Plotly. [n. d.]. Plotly for Python. ([n. d.]). Retrieved Jun 21, 2018 from https://plot.ly/python/

[6] Bogdan Vasilescu. [n. d.]. A data set for social diversity studies of GitHub teams. ([n. d.]). Retrieved Jun 7, 2018 from https://github.com/bvasiles/diversity

[7] B. Vasilescu, A. Serebrenik, and V. Filkov. 2015. A Data Set for Social Diversity Studies of GitHub Teams. In *2015 IEEE/ACM 12th Working Conference on Mining Software Repositories*. 514–517. https://doi.org/10.1109/MSR.2015.77